確率モデルに基づく自然言語文書からの知識抽出に関する研究

白井, 匡人 / SHIRAI, Masato

(開始ページ / Start Page)
1
(終了ページ / End Page)
106
(発行年 / Year)
2016-03-24
(学位授与番号 / Degree Number)
32675甲第375号
(学位授与年月日 / Date of Granted)
2016-03-24
(学位名 / Degree Name)
博士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)
(URL)
https://doi.org/10.15002/00013017

博士論文

確率モデルに基づく自然言語文書からの知識抽出に関する研究

STUDIES ON KNOWLEDGE EXTRACTION BASED ON PROBABILISTIC MODEL FROM NATURAL LANGUAGE DOCUMENTS

2016年3月

法政大学大学院理工学研究科
電気電子工学専攻博士後期課程

白井 匡人
Masato SHIRAI

目 次

第1章　序論

1.1　前書き

近年，インターネットの発達から大量のデータを容易に入手できるようになっている．これらの多種多様なデータは，様々な情報を含むが定型化されていないため，未整理のまま流れ去っている．多くのデータはテキスト形式で記述されている．特にストリーム中の文書は，タイムリな情報を扱うことから，極めて重要な情報源として注目されている．この大量のデータから知識を抽出するために，自然言語文書の解析手法が必要となる．自然言語文書の解析は，過去の様々な研究で扱われてきたが，コンピュータによる本質的な文書の理解には至っていない．自然言語文書の解析では，文脈の理解や談話の理解，多義語・同義語による文脈曖昧性・語義曖昧性の解消といった様々な問題が残されている．

文書の知識は，文書の持つ何らかの意味による記述を表す．定性的には，パターンやその組み合わせによる記述であり，言語モデルや論理モデルによって扱われる．定量的には，パラメータやパラメータモデルによる記述であり，統計モデルや確率モデルによって扱われる．また，文書の知識は，特定の知識を抽出するなら，知識獲得手法 (決定木やベイズ分類) を用いる．抽出した意味が不明であるとき，潜在意味抽出やクラスタリングを用いる．

文書から抽出できる表層的な情報として，単語や品詞，フレーズがある．単語の出現頻度や共起頻度は，文書の特徴となりえるが，高頻度語と重要語は異なる．熟語やコロケーションは一つのまとまりとして意味を成すため，個々の単語と区別する必要がある．また，表層的な情報は，語義や構文の曖昧さに強く影響を受ける．これにより，正確な意味を捉えることができない可能性がある．単語の字面を扱うだけでは構造を区別することができない．統計手法は表層的な情報しか扱えず，統計能力に限界がある．例えば，主成分分析や潜在意味分析を行うことにより，文書の持つ潜在的な意味を抽出する試みがなされている．しかし，抽出した語のまとまりを人手によって解釈する必要がある．本研究の狙いは，確率モデルによる文書集合の高度なモデル化の試みである．確率モデルは，確率分布により，文書集合全体を表現できる．確率モデルの利点としては，事前分布を用いることで文書に出現しない語を扱え，変数の依存を条件付き確率で表現できることが挙げられる．また，文書の単語分布は，多項分布に従うことが経験的に知られている．

過去の多くの研究でも，確率モデルによる文書に対するモデル化が行われている．ユニグラムモデルは，確率モデルに基づく文書集合のモデル化である．ユニグラムモデ

ルでは，全ての文書集合を一つの確率分布で表す．混合ユニグラムモデルでは，各クラスが確率分布を持つため，クラスごとの分布の異なりを表現できる．しかし，これらのモデルは文書が 1 つの確率分布に従うことを仮定する．このため，共通して表れる単語や，分野に依存してよく使用される単語といった，単語を出力する要因の混合を考慮することができない．本来単語が持つ意味は文書中の話題や文脈に依存して異なる．単語の意味の特定は，自然言語の持つ曖昧性によって定型化して抽出できないという点で困難である．自然言語文書から知識抽出を行うには，文書を特徴付ける要因の混合を考慮する必要がある．

本研究では，単語の意味や文書の内容を潜在状態を用いて確率的に捉え，単語の持つ意味を考慮した解析を行う．これによりクラスや話題が混合した文書集合から高水準な知識抽出を行う．確率モデルは，与えられた文書集合を観察値と見て推定される．しかし，動的な文書集合では，扱われる話題が時間と共に変動するため，特徴の変化が頻繁に起きる．文書ストリーム（ニュース記事等の新たな文書が逐次発生する文書集合）では，特徴の変動とデータ量が大量となることが知識抽出を困難にする要因となる．文書を特徴付ける要因の変化を捉えることがストリーム中での知識抽出において重要となる．また，ストリーム中では，変化に応じたモデルの修正が必要である．

本論文は，自然言語文書からの知識抽出を行うために 2 つの問題を論じる．第 1 の問題は，クラスや話題といった文書を特徴付ける要因の抽出である．文書の単語の分布には書き手やテーマ，ジャンルなど複数の要因による影響が混在していると仮定する．ここでは確率モデルを使用し，各単語の潜在状態を推定する．具体的には品詞分布の法則や潜在トピックを扱う．第 2 の問題は，動的な文書集合からの知識抽出である．ここでは，ストリーム中での各要因の特徴の変化が潜在トピックの変化として表現できることを仮定する．文書集合の変化に対応するため，ある間隔で到着する教師データを用いて新たに学習することにより，モデルを修正する．また，ストリーム中の文書には大きく 2 つの特徴がある．時間に影響されず各文書内で共通して表れる定常的な特徴と，文書中の話題の変化による特徴の変動である．このため，文書集合の特徴を表す分布は，1 つの定常状態に収束するのではなく，時間とともに変化することを仮定する．提案手法はオンライン学習を用いて特徴の変化に応じて潜在要因の変化を学習する．ここでは，事前分布の変化の学習と定常分布と変動分布の抽出を行う．本論文の論点は主に，(1) トピックモデルの適用範囲の拡大 (2) 文書ストリームへのトピックモデルの適用，という 2 点にある．

1.2 潜在要因による特徴抽出

本論文は，文書を特徴付ける要因として単語の潜在要因に着目する．ここでは，潜在要因として潜在トピックと品詞を用いる．この潜在要因は確率モデルにより抽出する．トピックモデルは，文書を潜在トピックの混合で表す確率モデルである．一般的に，トピックモデルは多重集合 (bag of words) 表現された文書を対象とする．多重集合は，文書中の単語の並びを考慮せず，各単語の頻度 $v_1, v_2, \cdots, v_n$ のみを表現する．

1.2.1　潜在トピックによる著者の特徴抽出

本研究では，文書が持つ複数の要因を抽出するために，トピックモデルにより著者の特徴を抽出する．文書はクラスが持つ複数の話題を含み，話題に関連して出現する単語分布が変化する．ここでの問題は，小説や社説などの文書には著者の特徴が表れるが，単語の分布の類似が著者によるものか，同一の話題によるものか判別することができない点にある．SF やミステリー等，同じジャンルを扱っていればジャンルに依存した単語が出現することが考えられる．同じ話題を扱っている文書には，共通の固有名詞などが多数使用される．このため，複数の要因を含む文書を扱うには，文書内の話題の混合を表現するモデル化が必要不可欠になる．提案手法は，トピックモデルを用いることで，文書を特徴付ける要因を捉える．トピックモデルでは，著者や新聞社といったクラスの特徴を話題の混合として表現する．

LDA(Latent Dirichlet Allocation) はトピックモデルの一種であり，1 つの文書が複数の潜在トピックの混合として表現される．トピックモデルでは，文書内の各単語に潜在状態であるトピックを仮定し，単語に潜在トピックを 1 対 1 で割り当てる．Author-Topic(AT) モデルは著者ごとにトピック分布を持つ点で異なる．AT モデルは複数の著者によって書かれた文書を扱えるが，本研究では単純化のため，各文書が単一の著者によって書かれている文書を扱う．このトピックモデルを用いることでクラスと各文書を潜在トピックの混合として表すことが可能となる．潜在トピックは，共通したクラスや文書に出現するといった何らかのまとまりを示す一種のクラスタである．潜在トピックが話題に対応すると仮定することで，話題の変化を潜在トピックの変化として表現する．しかし，小説や社説といった複数のジャンルを含む文書集合においても，適切なトピックを学習できるのか定かではない．

ここでは，トピックモデルを用い，各クラスを潜在トピックの確率分布として表現する．KL 情報量やカイ 2 乗値により各クラスとテスト文書のトピック分布を比較することで著者推定を行う．ここでトピック分布は小説と社説が混在した文書集合から求める．これにより，異なるジャンルの文書が混在した文書集合においても，著者ごとの分布の異なりを正しく捉えることを示す．

1.2.2　日本語の品詞分布特性によるジャンルの特徴抽出

文書が持つ複数の要因の抽出を行うために，ジャンルを特徴付ける要因を抽出する．文書ジャンルとは，文書の役割・状況を考慮した文書の種類であり，典型的には ”日記”, ”随筆”, ”小説”, ”社説”, ”報道記事”などがある．ここでは，ジャンルを特徴付ける要因として品詞に着目する．自然言語文書は，何らかの文法規則に基づき記述される．品詞は単語の働きに応じて種類分けされたものであり，文の意味を解釈するうえで有用である．

日本語文書の品詞分布に関しては，いくつかの先行研究によって品詞分布の法則が示されている [56][54]．樺島の研究では，名詞の割合によって他の品詞の割合が定まる

とし，樺島の近似式が示されている．大野らは，延べ語数，見出し語数のどちらにおいても日本語文書には品詞分布に法則性が存在することを指摘している．しかし，これらの先行研究で示されている近似式は品詞分布の特性を表してはいるが，実際の文書に適用すると当てはまりがあまり良くない．この原因としては，品詞の割合には文書のジャンルによって差がある [55] としながら，ジャンルごとの品詞分布の違いを考慮していないためであると考えられる．このため，品詞分布の法則がジャンルごとに成り立つと仮定する．品詞分布の法則がジャンルごとに適用可能ならば，品詞の割合だけからジャンル推定でき，極めて効率よく分類できる．

提案手法は，図 1.1 のようにジャンルごとに名詞の割合が変化すると仮定し，名詞の割合の事前分布にガウス分布を用いる．

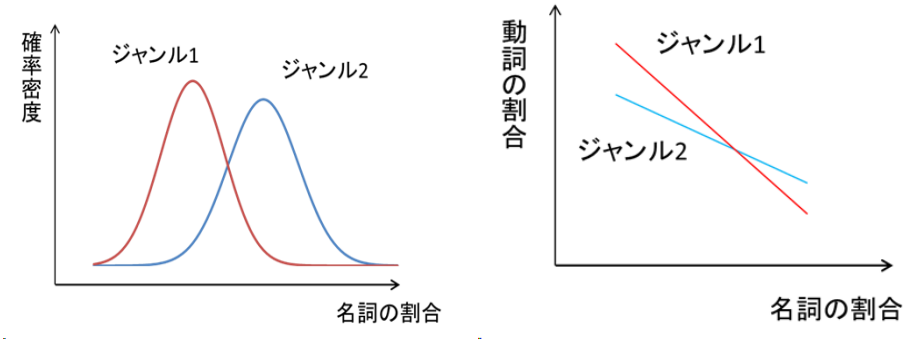

図 1.1: ガウス分布と線形回帰

1.2.3 検索語の意味抽出

文書を特徴付ける要因として，検索における語の働きを論じる．多クラス分類では，クラスを集合として与え，その特徴を基に各クラスに該当する文書を抽出する．これに対し，文書検索は数個の単語からなる検索語に対応する文書を抽出する．検索語は一般的に数個の単語からなり，この検索語との類似性により関連する文書を選択する．ここでの問題は，検索語の語義の曖昧性により，質問者の要望と異なる文書が選択される点にある．例えば，”APPLE”,”PIE”,”CHART” という 3 語からなる検索語から，質問者がアップルパイの売れ行きを知りたいのか，アップルコンピュータに関する円グラフを調べたいのかということを判断することは難しい．このため，検索語全体の意味を考慮して抽出する必要がある．

提案手法は，単語の係り受け関係を用いることで単語の依存構造を抽出する．さらに単語と係り受け語の組に対して潜在トピックを推定することで検索語の意味を考慮した検索を行う．

1.3　文書ストリームからの特徴抽出

本研究の目的は，潜在要因の変動を考慮することにより，動的な文書集合から特徴抽出を行うことにある．ニュースストリーム，マイクロブログに代表されるSNSデータは，極めて重要な情報源として注目されており，多数の人々が利用している．文書ストリームでは，一度新たな話題が発生すると関連した記事が短期間に多数発生し，単語分布が大きく変わるというバースト現象が発生する [18]．各話題に依存した単語分布は，新たなイベントの影響を強く受ける．例えば，オリンピックのような大きなイベントが起これば，そのイベントに注目が集まるため期間中はオリンピックに関する記事が急増する．期間が終わると異なる話題に遷移していきオリンピックに関する記事は減少する．

1.3.1　事前分布の学習による動的な特徴抽出

本研究では，文書ストリームからの特徴抽出を行うために事前分布の更新を行う．文書はクラスが持つ複数の話題を含み，話題に関連して出現する単語分布が変化する．ストリーム中では，この話題の特徴が大きく変化する可能性がある．この話題の変化によってクラスの特徴が変化していくため，動的な学習によりモデルを更新する必要がある．提案手法は，この特徴の変化を潜在トピックや単語を出力する事前分布の変化として捉える．

提案手法は，トピックモデルのオンライン学習により各クラスについてのディリクレ事前分布と各クラスの文書の特徴を学習する．トピックモデルのオンライン学習により事前分布を更新することで，クラス内での分布の変動を表す．バースト現象はクラスごとに生じ，変動が他クラスに影響を与えない．どのクラスでバーストが生じるかは多項分布確率的に生じると仮定する．従って複数のトピックモデルを混合して協調する上位モデルが必要である．クラスの出現確率とトピックの確率分布，単語の確率分布の事前分布をそれぞれ学習することで動的な文書集合からの特徴抽出を行う．

1.3.2　オンライン学習によるストリーム中の特徴抽出

文書ストリームからの特徴抽出を行うためにトピックモデルのオンライン学習を行う．これにより，ストリーム中の文書から定常的な特徴と変動的な特徴を抽出する．クラスを表す定常的な特徴は，クラス内の話題によらず一定して表れる．ストリーム中の話題の変化による特徴の変動は，時期やイベントに依存することから，現在の状態に合わせて特徴を学習する．この2つの特徴は，異なる性質を持つことから，それぞれ別の確率分布として用いることで高精度に文書集合をモデル化できる可能性がある．

提案モデルは，文書のクラスラベルを学習データとして使用することで，クラスごとのトピック分布を学習する．定常トピックと変動トピックという2種類の潜在トピックを用いることで，ストリーム中での潜在トピックの特徴の変化を捉える．定常トピッ

ク分布は，これまでに出現したすべて文書から得られるクラスの特徴を表し，変動トピック分布はウィンドウにより直近の文書の特徴を表す．ストリーム中では特徴の変動が不定期に起きる可能性があるため，直近の文書に対してウィンドウを遷移させていくことで特徴の変化を考慮する．

1.3.3　ストリーム中の複数のラベルを持つ文書からの特徴抽出

文書ストリーム中で複数ラベルを持つ文書から特徴抽出を行うためにトピックモデルを適用する．これにより，複数のラベルを持つマルチラベル文書から各ラベルの特徴を抽出する．マルチラベル文書中には話題の混合とラベルの混合による特徴が同時に出現する．このため，文書を特徴付ける要因を抽出することがより困難になる．また，ラベル間には”経済”と”市場”は共起しやすいが”芸術”と”健康”は共起しにくいといった依存性があることが考えられる．マルチラベル文書を扱うには，ラベル間の関係を考慮することが重要となる．文書を特徴付ける要因の混合と合わせてラベルの混合を扱うためのモデルの拡張が必要となる．

提案手法は，トピックモデルを用いることでマルチラベル文書をモデル化する．ここでは，ラベルの定常的な特徴とストリーム中に発生する局所的な変動を考慮する．各ラベルの特徴を学習し，ラベル間の共起関係を用いることで特徴抽出を行う．

1.4　公開論文

学術論文

1. Masato SHIRAI, Takashi YANAGISAWA, Takao MIURA，Context-based Query using Dependency Structures based on Latent Topic Model, Journal on Data Semantics (JoDS) Vol. 3, Issue 3 pp.157-168, ISSN: 1861-2032 (Paper), 2014.9

   テキストデータベース検索を高度化するために, 質問の文脈を的確にとらえる手法を提案する。ここでは潜在トピックモデルを用いて日本語文書の係り受け関係をモデル化し、検索に生かす方法を提案する。基本的なアイデアは、係り受け関係を文脈上で捉えトピックモデルにより高い確率のものを抽出することにある。

2. 白井 匡人, 三浦 孝夫，トピックモデルに基づくニュースストリームのオンライン分類，電子情報通信学会, Vol.J99-D，No.3, 2016.3

   トピックモデルに基づくニュースストリームのオンライン分類手法を提案する．ニュース記事のようなストリームデータでは新たな話題が逐次発生し，出現する話題が変化していくことから，一つの定常な確率分布によって分類を行うことは困難である．このことから，クラスが持つ特徴と話題による特徴の変化を加味して分類基準を変更する必要がある．提案手法では，クラスの定常的な特徴とストリーム中に発生する局所的な変動を考慮したトピックモデルによる分類を行う．

3. 白井 匡人, 三浦 孝夫，トピックモデルに基づく文書ストリームのマルチラベル分類，電子情報通信学会, Vol.J99-D，No.4, 2016.4

   文書ストリームを対象としたマルチラベル分類手法を提案する．特徴の変化が起こる文書ストリームでは新たな文書が逐次発生し文書集合が動的に変化することから，あらかじめ決まった定常な確率分布によって分類を行うことは困難である．このため，ラベルの特徴を動的に学習して分類を行う必要があり，マルチラベルでの特徴の変化も考慮することが求められる．提案手法では，ラベルの定常的な特徴とストリーム中に発生する局所的な変動を考慮したトピックモデルを用いる．各ラベルの特徴を学習し，ラベル間の共起関係をラベリングに利用することで文書ストリームのマルチラベル分類を行う．

国際会議論文（審査付き）

4. Masato SHIRAI, Takao MIURA, Online Classification with Partially Labelled Texts, International Conference on Web Intelligence (WI), a special session on Complex Methods for Data and Web Mining, 2014.8

   オンライントピックモデルを用いた文書ストリームの分類手法を提案する．文書ストリームではクラスの特徴が動的に変化するため，特徴の変化に応じて適応的に分類基準を変更する必要がある．また，文書中の話題が変化していくことから新たなクラスが出現する可能性がある．オンライントピックモデルによりクラスごとの動的学習を行い，ストリーム中の学習データから新たなクラスを獲得することで文書ストリームの分類を行う．

5. Masato SHIRAI, Takao MIURA, Document Classification Using POS Distribution, 16th East European Conference on Advances in Databases and Information Systems, 2012.9

   日本語文書の名詞に対する他の品詞の割合には相関関係があることが知られており、いくつかの近似式が示されている。しかしながら、すべての文書集合に同一の法則を適用する従来の手法では精度が悪い。そこで文書集合ごとにクラス分類を行うことで複数の近似式を得る。また、実験により本手法の有効性を示す。

6. Masato SHIRAI, Takao MIURA, On Domain Independence of Author Identification, 12th Intn'l Conf. on Intelligent Data Engineering and Automated Learning, 2011.9

   潜在的ディリクレ配分法 (LDA) では、１つの単語に付き１つのトピックを持ち、文書はトピックの確率分布で表される。同一の著者が書いた複数の文書を１つの文書と見なすことで、著者をトピックの確立分布で表せると考える。このトピックの確立分布を用いて著者推定を行う。また、実験によりその有効性を示す。

研究発表

7. 白井 匡人, 三浦 孝夫，時間変化を考慮した能動学習によるストリームのマルチラベル分類, 第7回データ工学と情報マネジメントに関するフォーラム (DEIM) , 2015.3

   能動学習を用いることでストリーム中で複数のラベルを持つ文書を対象としたオンライン分類手法を提案する．ニュース記事やマイクロブログのようなストリームデータでは，扱われる話題が変化することから，時間により文書の特徴や出現するラベルは大きく異なる．ラベル無し文書は新たな特徴を持つが，ノイズを含むため分類性能が悪化する可能性がある．提案手法では能動学習により，新たに到着した文書の特徴によって選別することで分類器の更新を行う．

8. 白井 匡人, 三浦 孝夫，トピックモデルに基づくニュースストリームのオンライン分類, 電子情報通信学会データ工学研究会（DE), 情報処理学会第159回データベースシステム研究会 第115回情報基礎とアクセス技術研究会 電子情報通信学会データ工学研究会合同研究発表会, 2014.8

   トピックモデルに基づくニュースストリームのオンライン分類手法を提案する。ニュース記事のようなストリームデータでは新たな話題が逐次発生し，出現する話題が変化していくことから，一つの定常な確率分布によって分類を行うことは困難である．このことから，クラスが持つ特徴と話題による特徴の変化を加味して分類基準を変更する必要がある．提案手法では，クラスの定常的な特徴とストリーム中に発生する局所的な変動を考慮したトピックモデルによる分類を行う．

9. 白井 匡人, 三浦 孝夫，トピックモデルに基づく文書ストリームのマルチラベル分類, 第6回データ工学と情報マネジメントに関するフォーラム（DEIM），2014.3

   文書ストリームを対象としたマルチラベル分類手法を提案する．文書ストリームでは新しい文書が逐次発生し文書集合が動的に変化することから，同一の分類基準を用いて分類を行うことは困難となる．このため，ラベルの特徴を動的に学習して分類を行う必要があるが，文書は一般に複数のラベルから構成されるため，マルチラベルでの特徴の変化も考慮することが求められる．提案手法では，トピックモデルによりラベルの特徴を動的に学習し，ラベルの分化・同化によりマルチラベルを構成する基底ラベルを自動的に変化させることで文書ストリームのマルチラベル分類を行う．

10. 白井 匡人, 三浦 孝夫，オンライントピックモデルによる文書ストリームの適応的分類, 電子情報通信学会データ工学研究会，2013.9

    オンライントピックモデルを用いた文書ストリームの分類手法を提案する．文書ストリームではクラスの特徴が動的に変化するため，特徴の変化に応じて適応的に分類基準を変更する必要がある．また，文書中の話題が変化していくことから

新たなクラスが出現する可能性がある．オンライントピックモデルによりクラスごとの動的学習を行い，ストリーム中の学習データから新たなクラスを獲得することで文書ストリームの分類を行う．

11. 白井 匡人, 島田 諭, 三浦 孝夫，トピックモデルのラベリングによる潜在的コミュニティの分析, 第5回データ工学と情報マネジメントに関するフォーラム (DEIM), 2013.3

    近年，トピックモデルを用いた研究が多方面で提案されている．例えば，ソーシャルネットワークのユーザ分析やコミュニティ抽出に用いる．しかしながら，ここで言うトピックとは，確率的基準による一種のクラスタであり，ユーザやコミュニティの意味を明示的に捉えることは自明ではない．本研究では，予め与えたラベルを用いて潜在的トピックをラベル付けし，明示的な意味を付与する手法を提案する．

12. 白井 匡人, 島田 諭, 三浦 孝夫，品詞分布を用いた日本語文書のジャンル分類, 第11 回 FIT（情報科学技術フォーラム），2012.9

    日本語文書の名詞に対する他の品詞の割合には相関関係があることが知られており，いくつかの近似式が示されている．しかしながら，すべての文書集合に同一の法則を適用する従来の手法では精度が悪い．そこで文書集合ごとにクラス分類を行うことで複数の近似式を得る．また，実験により本手法の有効性を示す.

13. 白井 匡人, 三浦 孝夫，確率モデルによる品詞分布の特徴推定, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2012.3

    日本語文書の名詞に対する他の品詞の割合には相関関係があることを利用し，時系列的に並んだ文書ストリームをクラス分類する手法を提案する．ストリームデータは標本化により近似し，統計的に正しい範囲で分類基準を随時訂正する手法を示す．また，実験により本手法の有効性を示す.

14. 白井 匡人, 三浦 孝夫，LDA を用いた著者推定, 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM), 2011.3

    潜在的ディリクレ配分法 (LDA) では，１つの単語に付き１つのトピックを持ち，文書はトピックの確率分布で表される．同一の著者が書いた複数の文書を１つの文書と見なすことで，著者をトピックの確率分布で表せると考える．このトピックの確率分布を用いて著者推定を行う．また，実験によりその有効性を示す.

第2章　研究背景と課題

2.1　文書における知識

文書を構成する文章テキストは何らかの規則に従っているが定型化されていないため，単語やフレーズ，品詞など何らかの単位に分割し，それらの出現頻度や共起関係などを抽出して解析を行う．文章中の単語は語の働きによって内容語と機能語に分けられる．内容語は名詞句など一般的な意味を成す語である．機能語は文節内で助詞など文法的な役割を果たす語である．機能語を用いた自然言語文書からの特徴抽出手法としては，一語当たりの平均文字数 [24], 読点による特徴付け [kim93]，n-gram 分布 matsuura99 がある．内容語による特徴抽出では，単語の重み付けにより，特徴語や重要語の抽出が行われている．TF*IDF は文書集合中の単語の重み付けの尺度として用いられる．TF*IDF は，単語の出現頻度 TF と逆文書頻度 IDF の積である．

$$TF = \frac{n_d^i}{N_d}\ ,\ \ IDF = log\frac{D}{df_i} \tag{2.1}$$

ここで，n_d^i は文書 d での単語 i の出現数，N_d は文書 d での総出現数，D は全文書数，df_i は単語 i を含む文書数である．TF*IDF では，クラスを特徴付ける固有名詞などが存在している場合に特に有効である．ここでの，問題は全ての単語が出現する文書の内容に依存せず，同一に扱われることにある．

また，文法上の特徴として日本語文書の品詞分布に対する法則性が示されている．樺島の法則は，延べ語数に関する品詞分布の特性である．ここでは，自立語を名詞，動詞，形容詞類（形容詞・副詞・連体詞），接続詞類（接続詞・感動詞）に分ける．名詞以外の品詞の割合は名詞の割合から求める．接続詞類を I としたとき各品詞の割合は以下の近似式で表される．

$$A_d = 45.67 - 0.60N \tag{2.2}$$

$$logI = 11.57 - 6.56logN \tag{2.3}$$

$$V = 100 - (N + A_d + I) \tag{2.4}$$

また，単語の異語数による品詞分布の特性として古典文学作品に焦点を当てた大野の語彙法則が提案されている．古典 9 作品において名詞の割合がもっとも高い文書を

左端に，名詞の割合がもっとも低い文書を右端におき二つの作品の品詞の割合を直線で結ぶと，他の作品の品詞の割合はこの直線状に垂直に並ぶ．この法則の定式化が水谷によってなされている [60]．ある 3 作品の名詞構成比を X_0，x，X_1 他の語類の構成比を Y_0，y，Y_1 としたとき以下の式が成り立つ．

$$y = Y_0 + \frac{Y_1 - Y_0}{X_1 - X_0}(x - X_0) \tag{2.5}$$

先行研究では，延べ語数，異語数に関して日本語文書には品詞分布に法則性があることを明らかにしている．

2.2 文書のモデル表現

文書のための確率モデルとしてユニグラムモデルが提案されている．ユニグラムモデルでは，全ての文書の単語が 1 つの確率分布 ϕ から出力されることを仮定する．ユニグラムモデルは，文書集合全体での高頻度語や低頻度語といった単語の頻度を表現できる混合ユニグラムモデルは，文書集合を k 個の確率分布 $\phi_1, \phi_2, \cdots, \phi_k$ で表す．単語分布 ϕ と各単語分布の出やすさ θ が与えられたときの文書 d の生成確率は以下の式で求まる．

$$P(d) = \sum_z P(z|\theta) \prod_{n=1}^{N_d} P(w_n|\phi_z) \tag{2.6}$$

ここで，N_d は文書 d での総単語数である．これによりクラスごとの出現頻度といった文書が生成される分布の異なりを表現できる．これらのモデルは各文書が 1 つの確率分布に従うことを仮定している．

文書を複数の確率分布の混合で表す手法として，確率的潜在意味索引付け (Probablistic Latent Semantic Indexing, pLSI) が提案されている [13]．pLSI は，トピックによる文書の次元圧縮手法であり，文書を単語次元からトピック次元に圧縮する．ここでは，文書ごとにトピックの混合比を学習する．これにより，文書をトピックの混合で表す．また，トピックは単語の混合で表される．文書 d で単語 w が生成される確率は以下の式で求まる．

$$P(w|d) = \sum_i P(w|z_i)P(z_i|d) \tag{2.7}$$

pLSI では，文書を特徴付ける複数の要因を考慮できる．しかし，トピックの混合比は与えられた文書集合に対して固定化している．また，pLSI は事前分布を持たないため，文書集合に存在しない単語を表現できない．このため，pLSI は新規の文書を扱えないという問題がある．

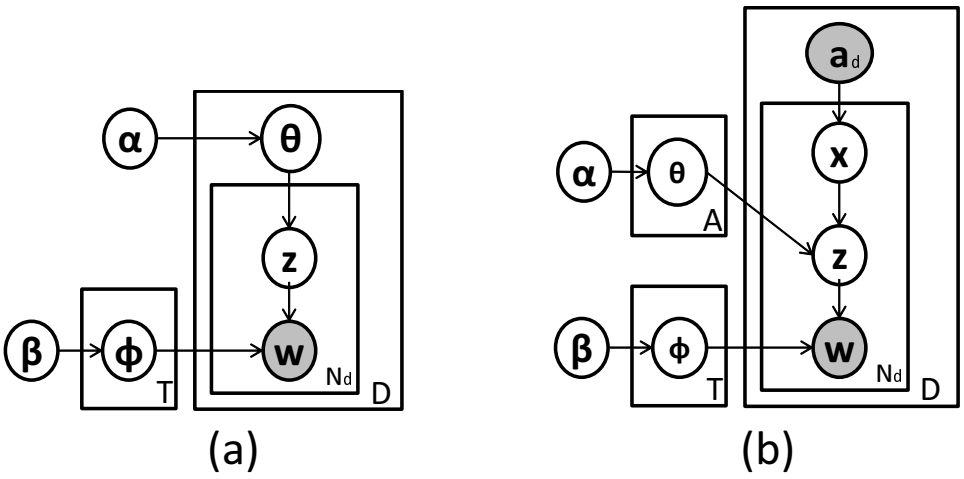

図 2.1: LDA と AT モデルのグラフィカルモデル

2.3 話題の抽出

トピックモデルは文書集合の分析に有効なモデルとして幅広く用いられている．トピックモデルでは, 文書を複数のトピックの混合として表現する．pLSI では, トピックの混合比を学習データの文書集合に依存して固定化していたが, LDA ではこの混合比を事前分布から生成する点で異なる．1 つの文書が 1 つのトピックで表される混合ユニグラムモデルに比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある．その中でも最近用いられているのが Latent Dirichlet Allocation(LDA) である [5].

LDA のパラメータ推定では，潜在変数であるトピック z，文書ごとのトピックの確率分布 θ，トピックの単語分布 ϕ を文書集合に対して尤度が最大となるように推定する．この潜在変数の推定には，ギブスサンプリング等が用いられる．ギブスサンプリングでは，サンプリングにより 1 つの単語に対して 1 つのトピックを割り当てる．この割り当ては，更新を行う単語以外のすべてのトピックの割り当てによって更新される．

図 2.1(a) では，LDA のグラフィカルモデルを示す. 図中の変数は, 図 2.1 左下に, ディリクレ事前分布 $Dir(\beta)$，図 2.1 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1 左上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N_d は各文書の単語数を表す．LDA の単語生成過程を以下で示す. まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する. 次に, 文書 d 内の i 番目の単語 w_i において，抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する.

文書 d の単語 w_i のトピックが z_i となる確率は，以下の式で求まる.

$$P(z_i|\mathbf{z}_{N\backslash i}, \mathbf{w}_N) \propto \frac{n^{w_i}_{z_i,N\backslash i} + \beta}{n^{(.)}_{z_i,N\backslash i} + V\beta} \frac{n^{d}_{z_i,N\backslash i} + \alpha}{n^{d}_{(.),N\backslash i} + T\alpha}$$

ここで，$\mathbf{z}_{N\backslash i} = \{z_1, ..., z_{i-1}, z_{i+1}, ..., z_N\}$ であり，i 番目の単語の割り当てを除外することを表す．$n^{w_i}_{z_i}$ は単語 w_i にトピック z_i が割り当てられた回数，$n^{(.)}_{z_i}$ は全単語中で z_i が割り当てられた合計である．$n^{d}_{z_i}$ は，文書 d で z_i が割り当てられた回数，$n^{d}_{(.)}$ は d にトピックが割り当てられた合計である．

LDA の拡張である AT モデルでは，図 2.1(b) で表される変数 x により文書の著者などの該当の文書が属するラベルの情報を持つ [31]．文書 d において $\mathbf{a}_d$ から著者 x は一様に得られ，トピック z は文書のトピック分布に基づきトピックから単語が生成される．LDA は文書ごとにトピック分布を持つが，AT モデルでは著者ごとにトピック分布を持つ点で異なる．また，AT モデルは文書に複数のラベルが存在することを仮定している．しかし，文書内でのラベルの偏りを考慮しておらず各ラベルは一様に生成される．文書 d の単語 w_i の著者が x_i，トピックが z_i となる確率は，以下の式で求まる．

$$P(z_i, x_i|\mathbf{z}_{N\backslash i}, \mathbf{w}_N, \mathbf{a}_d) \propto \frac{n^{w_i}_{z_i,N\backslash i} + \beta}{n^{(.)}_{z_i,N\backslash i} + V\beta} \frac{n^{x_i}_{z_i,N\backslash i} + \alpha}{n^{x_i}_{(.),N\backslash i} + T\alpha}$$

ここで，$n^{x_i}_{z_i}$ は z_i が著者 x_i に割り当てられた回数，$n^{x_i}_{(.)}$ は全トピックで x_i が割り当てられた合計である．

2.4 話題と文脈検索

近年多数の先行研究で情報検索においてトピックモデルが用いられている．Wei らは，アドホック検索に対して，LDA に基づく文書モデルを提案している [45]．ここでは，以下の式より，LDA を単語のスムージングに用いる．

$$P(w|D) = \lambda(\frac{N_d}{N_d + \mu} P_{ML}(w|d)) + \frac{\mu}{N_d + \mu} P_{ML}(w|D) + (1 - \lambda) P_{lda}(w|d) \qquad (2.8)$$

ここで N_d は文書 d の単語数である．μ はスムージングパラメータである．P_{ML} は，文書 d 内と全文書 D 内での単語 w の最尤推定量である．P_{lda} は，LDA による推定量である．Wei らの提案手法は，文書と単語の関係を捉えているが，検索語同士の関係を考慮していない．Yi らは，検索問題に LDA や PAM[20] といった異なるトピックモデルを適用したときの性能ついて調査している．これらの先行研究は既存のトピックモデルを検索に使用しており，検索を目的としたモデルを構築していない．このため，単語間の依存を捉えることができるが，検索語間の依存を捉えられない．

トピックモデルの問題点の一つは関連語や高頻度語の共起関係のモデル化の欠落にある．実際，どの文書にも出現する一般語といった高頻度語はモデルの性能を悪化さ

せる要因になる．ディリクレスムージングは文書と検索語の関係を捉えるのに有用であるが，検索語間の依存性を捉えるには不向きである．このため，検索問題に対応できるようにモデルの拡張が必要である．

2.5 文書ストリーム

本研究で扱う文書ストリームは，ラベル有り文書列 $\mathbf{L} = \{l_1, l_2, ..., l_{|L|}\}$ とラベル無しの文書列 $\mathbf{U} = \{u_1, u_2, ..., u_{|U|}\}$ の混合から構成されることとする．初期の学習文書集合から確率モデルを構築し，ストリーム中で新たに到着するラベル有り文書 l を用いて追加の学習を行う．文書ストリームからの知識抽出では，バースト現象やコンセプトドリフトを考慮した多数の研究がある [18][4][42][43].

一般的にトピックモデルはパラメータ学習を一括して行うため，文書集合の変化を想定していない．この問題を解決するためにトピックモデルのオンライン学習が提案されている [7][32][14]．トピックモデルのオンライン学習は，差分型ギブスサンプリングや EM アルゴリズムを用いることで，新たな文書に対してトピック分布を推定できる．Canini らの差分型ギブスサンプリングでは，現在までに得られたパラメータを基に各単語にトピックを割り当て，その後活性化ステップにおいて過去のトピック割り当てをランダムに選択し，再サンプリングを行う．

LDA のストリームへの対応として，Banerjee らによるオンライン学習手法が示されている [2]．このオンライン学習手法では，まず現在までの文書集合に対して，パラメータを推定する．得られたパラメータを基に新たな文書のパラメータを MAP 推定する．新たな文書 d の単語 w_i のトピック割り当ては以下の式で推定する．

$$P(z_i|\mathbf{z}_{i-1}, \mathbf{w}_i) \propto \frac{n^{w_i}_{z_i,i\backslash i} + \beta}{n^{(.)}_{z_i,i\backslash i} + V\beta} \frac{n^{d}_{z_i,i\backslash i} + \alpha}{n^{d}_{(.),i\backslash i} + T\alpha} \tag{2.9}$$

一括学習では単語 w_i のトピックを $\mathbf{z}_{i,N\backslash i}$ より推定し，単語列 $\mathbf{w}_N$ 内で繰り返し学習するが，オンライン学習では，現在までの観測データ $\mathbf{z}_{i-1}$ を用いて w_i のトピックを推定する点で異なる．ここでは，過去の文書のトピックに関して再推定は行わない．

2.6 関連研究と課題

2.6.1 複数のラベルを持つ文書からの特徴抽出

文書に複数の話題が含まれるのと同時に，文書が複数のラベルを持つ場合がある．複数ラベルを持つ文書からの特徴抽出手法として，McCallum ら [23] は，ガウス分布を前提とした EM アルゴリズムを用いる方式を提案している．ラベルの組合わせの個々に対して，各確率分布パラメータと混合比を計算し，最大尤度のマルチラベルを割り当てる．しかし，ラベルの組み合わせ数は $2^n - 1$ であるため，ラベル数の増加に伴い

計算時間が膨大となる．ラベル数が多い文書集合では，ラベルの組み合わせに対して特徴を抽出することは困難となる．複数ラベルを持つ文書から個々のラベルの特徴を抽出する手法として，Xioufis らは，ラベルごとに当該のラベルに属する文書と，属さない文書をそれぞれ別のウィンドウに保持することで，ラベルの特徴を抽出する手法を提案している [46]．この手法は，文書ストリームにも対応しており，新たな学習データが到着するたびに，最も古い文書と到着した文書が入れ替えウィンドウを更新する．ウィンドウによって各ラベルの出現の割合に影響されず，直近に出現した文書の特徴も加味することができる．また，ストリーム中で効率的な学習を行う手法として，能動学習を用いた分類手法が提案されている [43]．Wang らは，新たに到着した文書チャンク内で能動学習を行い，分類器の構築に有効なラベル無し文書を抽出する．ラベル有り文書とラベル無し文書を共に学習に用いることで各ラベルの特徴を表す．本研究では，複数ラベルを持つ文書からの知識抽出を行うためにトピックモデルを適用する．

2.6.2　トピックモデルによる知識抽出

トピックモデルの利点は，対象の文書集合に応じて潜在変数を仮定することで，目的に合ったモデル化が行える点にある．先行研究では，様々な文書集合に対してトピックモデルを適用している．Ida らは，オンラインレビューのような数値レートを持つ文書集合に対するモデル化として，領域依存と領域独立のトピックを用いた Domain-Dependent/Independent Topic Switching Model を提案している [38]．DDITSM の領域依存トピックは Labeled LDA と同様にクラスとトピックが 1 対 1 で対応しており，さらに文書の数値レートを用いてクラスを分けることで極性情報を取り入れることができる．領域独立のトピックはディリクレ過程を用いることであらかじめ決められたトピック数ではなく，データ集合に応じたトピック数を推定することができる．DDITSM では，文書のクラスを教師情報として持つため，領域依存トピックとしてクラスの特徴を抽出することが可能であるが，領域依存トピックの時間変化を考慮していない．Kawamae らは，タイムスタンプが与えられた文書集合に対して，定常トピックと時制トピックを検出する Theme Chronicle Model を提案している [17]．TCM では，スイッチ変数により各単語を出力するトピックを選択することで，話題，トレンド，文書依存，バックグラウンドという 4 種類のトピックを学習する．話題は文書ごとにトレンドの分布を持ち，タイムスタンプによる時刻によってトレンドの出現が変化する．TCM では，TOT と同様に教師情報としてタイムスタンプを持つことで，トピックの特徴が時間変化する文書集合を扱える．しかしながら，ここではクラスのトピック分布やその変化を直接的には考慮していないまた，トピックモデルの拡張として，トピックに階層構造を取り入れたモデルが提案されている [20]．Li らによる Pachinco Allocation モデルでは，有向非循環グラフを用いてトピック構造を推定することで，トピック同士の関係を捉えることができる．トピックに階層構造を取り入れることでより高精度に文書集合のモデル化を行える可能性があるが，ストリーム中では階層構造も変化していくことが考えられるので本稿ではトピックの階層構造を扱わない．

2.6.3 トピックモデルの時系列データへの適用

トピックモデルでは，時系列データへの適用手法が提案されている．Blei らは，各時刻にトピックモデルを仮定し，その事前分布の依存をモデル化した Dynamic topic model(DTM) を提案している [6]．ここでは，時間単位のトピックの変化を捉えている．これに対し，時間情報を文書に組み込んだ時系列モデルとして Wang らによる Topics over Time モデルがある．ここでは，各文書が時間情報 t を持つことで，時間ごとのトピックの出現の変化を表現する．さらに，TOT モデルに著者情報を組み込んだモデルとして Naveed らによる Author-Topic-Time(ATT) モデルがある [27]．このモデルではトピックと著者に時間情報を与え，時間によるトピックと著者の出現の変化を考慮している．これにより，文書及び時期によるトピック分布と著者の分布を捉える．これらの時系列モデルは，既知の時系列データに対するモデル化であり，新たに文書が発生する動的な文書集合を対象としていない．

動的な文書集合への適用として，Iwata らによる Online Multiscale Dynamic Topic モデルはオンライン学習可能な DTM の発展形である [39]．このモデルは，異なるスケールでの単語の依存性を用い，複数のスケールでトピックの時間発展を捉える．AlSumait らは時刻 t でのトピックごとの単語分布に時刻 $t-1$ での出現確率を事前分布として用いることで，それ以前のデータを必要としない LDA のオンライン学習手法である OLDA を提案している [1]．

2.6.4 課題

本研究では，確率モデルに基づく文書集合のモデル化により自然言語文書から知識抽出を行う．特に，本研究はトピックモデルのオンライン学習により，文書ストリームからの特徴抽出について取り組む．ここでは，季節や月などの一定期間ごとに変化する特徴を捉えるのではなく，不定期に変化する特徴を捉える適応的な特徴抽出手法を提案する．

第3章　潜在トピックによる著者の特徴抽出

3.1　前書き

近年，文書の特徴を抽出するために様々な手法が用いられ，文書の特徴によって著者推定が行われている．著者推定に用いられる手法・特徴としては，出現する語のなんらかの特徴を捉えることが一般的である．構文解析後, 文節では名詞句など一般的意味を成す語は内容語，文節内で助詞など文法的な役割を果たす語は機能語と呼ぶ．機能語に関しては一語当たりの平均文字数 [24], 読点による特徴づけ [6]，n-gram 分布 [7] などが挙げられる．内容語の分布は著者の特徴となるが，単語の分布の類似が著者によるものか，同一の話題によるものか判別することは困難である．

本研究では潜在的ディレクレ割当て (Latent Dirichlet Allocation, 以下 LDA)[5] を使うことにより，単語の潜在的トピックを推定する．これにより，文書を複数のトピック分布で表し，トピックの確率分布により著者推定を行う．LDA では，文書が 1 つのトピックに対応するのではなく，文書内に出現する各単語にトピックが確率的に対応すると見なす．この結果，各文書に複数のトピック分布が対応する．逆に，トピックの確率分布は文書ごとに存在するが，複数トピックをまとめて 1 つの文書に対応させ，著者にトピックの確率分布が対応すると考える．この仮説を著者トピックモデル (Author-Topic モデル)[31] という．著者が持つトピックの確率分布は著者の特徴を表したものである．文書が持つトピックの確率分布にはその文書の著者の特徴が表れる．LDA を用いれば，著者の単語の分布とテスト文書の単語の分布を比較することにより， トピックの確率分布によって著者推定が行える．本研究では，著者のトピックの確率分布とテスト文書のトピックの確率分布を (KL 情報量やカイ 2 乗値等を用いて) 比較する．

小説では SF やミステリー等，ジャンルは同じであっても他の著者と厳密に同じ話題について書かれている作品はほとんどない．社説のような他の新聞社と同じ話題について論じられることが多い文書を用いることによって，例え話題が同じであっても新聞社ごとに表れる新聞社の特徴によって著者推定が行えることを示す．著者推定は著者の特徴によって推定を行うため，著者ごとに特徴があることが前提であり，著者の特徴量によって推定精度が変化すると考えられる．著者同士のトピックの確率分布を比較することで，他の著者との分布の異なりが大きい著者の推定精度が高くなると考えられるため実験で確認する．

第 3.2 章では LDA について述べ，第 3.3 章では推定手法について述べる．第 3.4 章で

は実験により有効性を示し，第 3.5 章で結びとする．

3.2 LDA による著者推定

3.2.1 トピックモデル

トピックモデルとは, 1 つの文書が複数のトピックの混合として表現されるという仮定である．1 つの文書が 1 つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある. その中でも最近用いられているのが LDA である. 確率的潜在意味索引付け (Probabilistic Latent Semantic Indexing, pLSI) は, LDA と違って, トピックと単語の多項分布にそれぞれディリクレ事前分布を導入し, 混合比を学習データの文書集合に依存して固定化している．一方, LDA ではこの混合比は事前分布から動的に生成する点で異なる. LDA は学習データだけに依存しないが, 代わりに特定の確率モデルを仮定するため, 柔軟な観点で文書を扱える可能性がある.

2 章図 2.1(a) は LDA のグラフィカルモデルである．図中の変数は, 図 2.1 左に, ディリクレ事前分布 $Dir(\beta)$, 図 2.1 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1 上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N は各文書の単語数を表す. LDA の単語生成過程を以下で示す. まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する. 次に, 文書 d 内の i 番目の単語 w_i において, 抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し, そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する.

3.2.2 著者トピックモデル

トピックモデルでは，著者は各文書にある様々な潜在的トピックについていくつかの文書を書くため，明示的な著者の情報は与えられない．そこで，トピックモデルに類似した著者トピックモデルでは，2 章図 2.1(b) で表される新しい変数 x により著者の情報を持つ．文書 d において $\mathbf{a_d}$ から著者 x は一様に得られ，トピック z は文書のトピック分布に基づき確率的に選択され, 選択されたトピックから単語が生成される．ここで各文書はディリクレ事前分布から得られるトピックの分布と関連していると仮定する．このため，潜在的トピック z の多項分布から得られる単語 w はトピックについてのディリクレ分布 $p(\phi)$ から得られるトピック z と関連している．

著者トピックモデルでは文書が複数の著者によって書かれることを考慮しており，トピックモデルは著者が一人のときの著者トピックモデルと見なすことができる．

3.2.3 パープレキシティ

まず, モデルの評価として, モデル内での Topic, 繰り返し回数を決定する. そこで言語モデルの性能評価には一般に, テストセット・パープレキシティと呼ばれる情報理論に基づく客観的評価手法を用いる. テストセット・パープレキシティは次式で計算される.

$$PP = 2^{H(p)}, H(p) = -\frac{\sum_X \log_2 p(X)}{N}$$

ここで $X_1 \cdots X_N$ は評価用のテキスト集合とする． 各語の確率をもとめ，そのすべての情報量の平均をＥＸＰの係数としている．すべての語の生起確率を求めるには，トピックの確率×語の生起確率を用いる．P_M は言語モデル M による $X_1 \cdots X_N$ の生成確率を表す. パープレキシティ値が高いほど単語の特定が難しく, 言語として複雑である. よって, パープレキシティの値が低いほど言語モデルの性能が高いと評価できる.

$$P_M(X_1 \cdots X_N) = \Pi_{i=1}^{N} p(X_i)$$

$$PP(s) = P_M(X_1 \cdots X_N)^{-\frac{1}{N}} = 1/\sqrt[N]{p(X_1) \cdots p(X_N)}$$

3.2.4 LDA のパラメタ推定

LDA のパラメタを推定する方法としては，EM アルゴリズムやギブスサンプリング等が挙げられるが，EM アルゴリズムでは局所最適解に陥りやすいという欠点があるため，本研究ではパラメタ推定にギブスサンプリングを用いる．ギブスサンプリングを用いてトピックの確率分布を更新することによって，各単語につくトピックが変化する．トピック j の確率分布は，トピック j 以外のすべてのトピックの確率分布によって更新される．これをすべてのトピックに対して行い，更新を繰り返すことにより，尤もらしい ϕ と θ の値が推定される．ある文書内では，ディリクレ分布によってトピックの確率分布には偏りがでるため，トピック内には，同じ文書で出現する単語が集まりやすくなっている．ここでギブスサンプリングの更新式を以下の式で定義する．

$$P(z_i = j | w_i = m, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta} \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha} \tag{3.1}$$

また，ギブスサンプリングの結果，推定される ϕ と θ の値は以下の式で求める．

$$\phi_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta} \tag{3.2}$$

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha} \tag{3.3}$$

C_{mj}^{WT} は単語 m がトピック j に割り当てられた回数，C_{dj}^{DT} は文書 d がトピック j に割り当てられた回数，V は全単語数，T は全トピック数である．

3.3 著者推定

著者が既知であり同じ著者が書いた文書を 1 つの文書と見なした文書の集合を Y，著者が未知である文書の集合を Z とする．この文書の集合である Y と Z を同時に LDA にかけることによって，著者ごとのトピックの確率分布 θ とテスト文書のトピックの確率分布 θ とトピックごとの語の確率分布 ϕ を得る．ここで得られたトピックの確率分布 θ は，各文書ごとに 1 つであるが，これらはトピックごとの単語の確率分布である ϕ に影響を受ける．ここでは各著者はトピック分布で特徴つけられると仮定しており，テスト著者のトピック分布と類似しているものを探せばよい．ある著者のトピックの確率分布 θ とある著者が正解であるテスト文書のトピックの確率分布 θ が類似し，トピックの確率分布は著者の特徴を表すものとなり，テスト文書のトピックの確率分布 θ は正解である著者のトピック確率分布 θ と最も類似すると考えられる．よって，本研究では，2 つのトピックの確率分布を比較することにより著者推定を行う．トピックの確率分布を比較する方法として，2 つの確率分布の差の尺度である KL 情報量と 2 つの分布がどのくらい独立しているかを調べるカイ 2 乗値を用いる．θ_x を著者 x のトピックの確率分布，θ_d をテスト文書 d のトピックの確率分布として，トピックの確率分布の KL 情報量を次のように定義する．

$$KL(\theta_x//\theta_d) = \sum_i \theta_{xi} log \frac{\theta_{xi}}{\theta_{di}}$$

また，トピックの確率分布の X^2 値を次のように定義する．

$$X^2 = \sum_i \frac{(\theta_{di} - \theta_{xi})^2}{\theta_{xi}}$$

3.4 実験

3.4.1 実験準備

実験には 2 つのコーパスを用いる．1 つ目のコーパスには，夏目漱石，森鴎外，島崎藤村の 3 人の著者の小説を 10 作品ずつ計 30 作品 (表 4.1) を用いる． 各小説は，1 章と 2 章を学習データとし，3 章以降をテストデータとする．2 つ目のコーパスには，2007 年度の毎日新聞，朝日新聞，読売新聞の社説を用いる．各社説は，9ヵ月分を学習データ，3ヵ月分をテストデータとする．社説の総数を表 3.2 に示す．実験データには MeCab による形態素解析を行い，名詞，動詞，形容詞，副詞のみを抽出して用いる．また，各実験データ中で 1 度しか出現しない単語は取り除いて用いる．

これらのコーパスを用いて 4 つの実験を行う．まず最初に第 1 の実験では，パープレキシティを測定することによりモデルの評価を行う．次に第 2 の実験では，小説と社説を用いて著者推定を行う．第 3 の実験では，小説の 3 人の著者の内，夏目漱石，森

表 3.1: 小説リスト

著者	作品名
夏目 漱石	坊っちゃん, 硝子戸の中, 彼岸過迄, こころ, 草枕, 行人, 道草, 門, 野分, 三四郎
森 鴎外	かのように, 魚玄機, 雁, 最後の一句, 細木香以, 心中, 二人の友, 高瀬舟, 寒山拾得, 安井夫人
島崎 藤村	嵐, 旧主人, 食堂, 岩石の間, 千曲川のスケッチ, 船, 分配, 藁草履, 家 (上), 刺繍

表 3.2: 社説の総数

	毎日	朝日	読売
学習データ	515	507	513
テストデータ	177	171	175

鴎外のみを学習して 3 人の著者を推定する．そして最後に，第 4 の実験では小説と社説を混合して著者推定を行う．

パラメタ推定には乱数を使用しているため，推定結果は乱数によって結果が異なる．このため，乱数の初期値を変化させて実験を 10 回行い，その平均値を正解率とする．また，ディリクレ分布のパラメタである α，β の値は 0.01 とし，トピック数は小説で 100，社説で 80，小説と社説の混合で 100 とする．ギブスサンプリングの繰り返し回数は 500 回とする．

3.4.2 評価方法

実験の評価方法としてパープレキシティと F 値を用いる．F 値は再現率と精度の調和平均であり，実際に正であるもののうち，正であると予測されたものの割合である再現率を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \tag{3.4}$$

また，正と予測したデータのうち，実際に正であるものの割合である精度を次のように定義する．

$$P_i = \frac{a_i}{a_i + b_i} \tag{3.5}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この 2 つの式の調和平均である F 値は次のように定義する

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad F = Average_i F_i$$

3.4.3 実験結果

まず初めに，小説，社説，小説と社説の混合のテストセットパープレキシティの値を表 3.3 に示す．ここでトピック数は 100，ギブスサンプリングの繰り返し回数は 500 回である．社説のパープレキシティの値は小説の 1.266 倍となっており，もっとも悪い値を示している．

表 3.3: パープレキシティ（トピック数 100）

小説	社説	小説＋社説
1140	1443 (1.266)	1393 (1.222)

著者推定を行った結果の正解率を表 4，表 5 に示す．小説での著者推定の全体の正解率は，KL 情報量で 97%，X^2 値で 92% である．社説では，著者推定の全体の正解率は，KL 情報量で 71%，X^2 値で 59% である．

表 3.4: 正解率（小説）

	夏目	森	島崎	全体
(KL)				
再現率	100	94	97	97
精度	92	100	100	97
F 値	96	97	98	97
(X^2)				
再現率	75	100	97	91
精度	100	79	100	93
F 値	85	88	98	92

著者同士のトピックの確率分布の比較結果を表 3.6 と表 3.7 に示す．著者推定の正解率の高くなる小説では著者同士の差が大きくなっていることが確認できる．また，KL 情報量による比較で最も F 値の小さい毎日新聞では他の新聞社との差が小さくなっている．

そして次に，各著者で確率の高いトピックのトップ 10 を表 3.8 と表 3.9 に示す．表 3.8 を見ると，すべての著者に共通して表れているトピックはトピック 84 だけである．

表 3.5: 正解率（社説）

	毎日	朝日	読売	全体
(KL)				
再現率	57	89	66	71
精度	67	71	75	71
F 値	61	79	69	71
(X^2)				
再現率	65	63	42	57
精度	48	64	71	61
F 値	55	62	51	59

表 3.6: 著者同士の比較 (小説)

	夏目	森	島崎
(KL)			
夏目	0	5.837	7.558
森	8.045	0	7.486
島崎	8.547	6.296	0
(X^2)			
夏目	0	128372	110690
森	15285	0	32309
島崎	68433	112637	0

表 3.9 では，すべての著者で確率の高くなっているトピックはトピック 76，70，53 となり，共通して確率の高くなるトピックが多い．

第 3 の実験では，夏目漱石と森鴎外のみを学習したことから，夏目漱石でも森鴎外でもなければ島崎藤村と判断する．このため，閾値を設けその値を $KL = 5$，$X^2 = 30000$ とする．著者推定の結果を表 3.10 に示す．正解率は，KL 情報量で 94%，X^2 値で 91% である．

最後の実験では小説の 3 人の著者と社説の 3 つ新聞社の計 6 人の著者で著者推定を行う．その結果を表 3.11 に示す．正解率は，小説では KL 情報量で 95%，X^2 値で 86%，社説では KL 情報量で 67%，X^2 値で 51% である．

ここで 6 人の著者同士の比較結果を表 3.12 に示す. 表 3.12 より小説の著者と社説の著者では，大きく異なっていることが見てとれる．次に各著者で確率の高くなるトピックのトップ 10 を表 3.13 に示す．小説で共通して確率の高くなるトピックはトピック 12 だけであるが，社説ではトピック 37，40，66，75，79 と 5 つのトピックが共通して表れている．

表 3.7: 著者同士の比較 (社説)

	毎日	朝日	読売
(KL)			
毎日	0	1.095	1.005
朝日	0.655	0	1.843
読売	0.726	1.410	0
(X^2)			
毎日	0	3276	29486
朝日	26959	0	41355
読売	21002	58470	0

表 3.8: トピックの確率のトップ 10（小説）

夏目		森		島崎	
トピック	確率	トピック	確率	トピック	確率
17	0.1015	34	0.1526	78	0.0931
30	0.0607	28	0.1036	84	0.0905
84	0.0502	45	0.0925	34	0.0864
95	0.0497	76	0.0705	61	0.0741
91	0.0492	7	0.0541	60	0.0608
54	0.0487	84	0.0541	53	0.0592
26	0.0454	17	0.0538	14	0.0588
41	0.0434	1	0.0417	62	0.0498
76	0.0336	55	0.0393	19	0.0353
31	0.0279	96	0.0364	5	0.0330

3.4.4 考察

著者推定の正解率は，KL 情報量による比較で小説のときの 97%が最も高い値となり，小説と社説の混合での 67%が最も低い値となっている．また，パープレキシティ値は小さい値となっている．

高い正解率となる小説での著者推定では，表 6，表 7 より各著者に明確な相違ができていることがわかる．また，表 8 より，トピック 84 だけがすべての著者に共通して表れていることがわかる．ここでトピック 84 の内訳を表 3.14 に示す．表 3.14 より，トピック 84 には特徴的な語が出現していない．

一方，社説では著者同士の相違が少なくなっている．特に毎日新聞と朝日新聞では，KL 情報量の値が非常に小さい値となっており，毎日新聞から朝日新聞を区別することは困難となっている．これは，著者に共通して表れるトピックが多くなっているために F 値が減少していると考えられる．しかしながら，社説であっても著者の個性によって著者推定が行えている．

表 3.9: トピックの確率のトップ 10（社説）

毎日		朝日		読売	
トピック	確率	トピック	確率	トピック	確率
76	0.0692	29	0.0727	76	0.0724
75	0.0522	27	0.0643	54	0.0594
13	0.0465	13	0.0483	75	0.0520
40	0.0437	70	0.0427	26	0.0451
27	0.0428	53	0.0410	70	0.0382
70	0.0420	36	0.0384	72	0.0375
64	0.0415	76	0.0363	5	0.0340
72	0.0352	46	0.0342	40	0.0338
10	0.0334	47	0.0331	7	0.0331
53	0.0290	14	0.0306	53	0.0321

表 3.10: 正解率（小説）

	夏目	森	島崎	全体
(KL)				
再現率	97	91	92	93
精度	88	100	95	94
F 値	92	95	93	94
(X^2)				
再現率	74	100	96	90
精度	100	82	94	92
F 値	85	90	95	91

KL 情報量は情報の量であるのに対し，X^2 値は誤差の量である．このため，得られる F 値は異なるが，小説や社説の著者推定では最も F 値が高くなる著者は同じになっている．表 6 より小説の著者間には明確な特徴が表れていることから，著者トピックモデルにより著者に明確な特徴が与えられ，著者推定が行えていると考えられる．

小説と社説の混合のテストセットパープレキシティは社説の値より小さい値となっており，正解率は KL 情報量による比較で小説が 95%，社説が 67%である．また，表 3.12 より，小説と社説には明確な相違があることがわかる．表 3.13 では小説と社説で共通して表れるトピックがなく，2 つの領域を正確に分離できていると考えられる．

これらの結果から，領域に依存するトピックを検出し，これらが共通するものではないことから，著者推定が領域に独立して動作すると考えられる．

表 3.11: 正解率（混合）

	夏目	森	島崎	小説	毎日	朝日	読売	社説
(KL)								
再現率	100	93	92	95	51	81	69	67
精度	87	100	100	96	62	74	65	67
F 値	93	96	96	95	56	77	66	67
(X^2)								
再現率	88	100	92	93	41	61	50	50
精度	92	48	100	80	48	55	54	53
F 値	89	63	96	86	43	57	51	51

表 3.12: 著者同士の比較（混合）

	夏目	森	島崎	毎日	朝日	読売
(KL)						
夏目	0	3.256	3.927	13.018	11.373	13.606
森	5.274	0	4.653	13.442	12.591	13.688
島崎	5.879	4.838	0	14.096	12.996	13.019
毎日	10.133	10.471	11.374	0	0.620	0.954
朝日	10.615	10.209	11.282	0.467	0	3.173
読売	9.804	10.599	11.547	0.338	0.654	0
(X^2)						
夏目	0	73341	56024	86067	101628	75854
森	6774	0	23866	54325	46076	67472
島崎	17809	41944	0	137839	118245	171731
毎日	851838	1231037	1886942	0	141	6
朝日	834555	1252465	1636291	16485	0	18
読売	1081328	1279498	860258	27994	322294	0

3.5 結び

本論文では，LDA の構造内で著者推定の領域独立を示した．すなわち，トピックが著者間の潜在的な区別できる特性を捉えることから，著者トピックモデルは領域カテゴリーから独立していることを経験的に示した．実験を通して，著者トピックモデルは (1) 各著者に対して明確な単語の分布を表すいくつかのトピックを検出する，(2) 一般的でない文書でもトピックは対象のカテゴリーに依存する，(3) これらのトピックの混合の意味によって動作する，ということが言える．

表 3.13: トピックの確率のトップ 10（混合）

夏目		森		島崎		毎日		朝日		読売	
トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率
32	0.1395	62	0.1300	87	0.2096	37	0.1398	43	0.0936	37	0.1873
24	0.0982	32	0.1158	9	0.1117	36	0.0958	92	0.0918	46	0.0827
34	0.0792	70	0.1096	52	0.0972	17	0.0705	37	0.0849	40	0.0765
9	0.0668	49	0.0925	67	0.0792	40	0.0675	75	0.0576	36	0.0652
4	0.0569	12	0.0711	98	0.0637	75	0.0559	40	0.0522	26	0.0584
91	0.0523	65	0.0625	12	0.0543	79	0.0481	66	0.0515	99	0.0450
96	0.0521	4	0.0543	27	0.0515	26	0.0416	79	0.0514	66	0.0429
31	0.0514	45	0.0529	100	0.0454	46	0.0412	17	0.0483	79	0.0402
100	0.0426	24	0.0464	25	0.0442	66	0.0401	36	0.0458	75	0.0331
12	0.0402	31	0.0306	71	0.0407	97	0.0297	15	0.0329	84	0.0287

表 3.14: トピック 84（小説）

単語	確率
よう	0.1126
いる	0.1085
時	0.0398
もの	0.0371
ない	0.0361
そう	0.0338
思う	0.0332
学校	0.0227
くれる	0.0211
一	0.0179

第4章　日本語の品詞分布特性によるジャンルの特徴抽出

4.1　前書き

計量言語学は，言語現象を統計手法を用いて定量的に分析する分野である．書き言葉である文章テキストは言語現象の例であり，計量文体学や計量的コーパス言語学などは計量言語学として盛んに論じられている．

文章は，何らかの文字列が一定の文法規則に基づいた文の集合体であり，定型化されていないデータとして典型的である．記号列が何らかの規則に従ってはいるが，定型化されていないため，単語やフレーズ，品詞など何らかの単位に分割し，それらの出現頻度や共起関係・同時出現などを抽出して，定量的に解析できる．この手法は総称してテキストマイニングと呼ばれている．計量文体分析の分野では一世紀以上も前から，文章を構成する要素の特徴を定量的に分析する方法で文章の執筆者の推定などを行っている．例えば，一般人が書いた短い文章の書き手の同定・推定では短い文章でも書き手がほぼ同定できる [58]．このような手法は，手紙や脅迫文の著者同定，スパムメール識別，ブログ分類にも応用されている．もっと一般的は応用分野として，アンケート分析などを用いた顧客サービスの自動評価，テキストマネジメントと情報・知識の抽出などがある．

文書は，その目的や対象とする読者のためにある統一的な難易度で作成され，例えば“増える”という日常表現は，論文などの公式的な分野では“増加する”という表現に代わって利用される．このように，文書ジャンルとは，文書の役割・状況を考慮した文書の種類であり，典型的には”日記”, ”随筆”, ”小説”, ”社説”, ”報道記事”などがある．

本研究では延べ語数での品詞分布を用いることで文書のジャンル類別が行えることを示す．品詞分布を特徴量として用いてジャンル分類が高精度に実施可能ならば，次元数が品詞の種類数となり，小さい次元で文書の比較を行うことができる．この結果，次元数が極めて少ないことでデータ数が増えても必要なメモリ量は少量で済むという利点がある．

第 4.2 章ではジャンル分類について述べ，第 4.3 章は日本語文書の品詞分布特性を論じる．第 4.4 章では提案手法について述べ，第 4.5 章では実験によりこの有効性を示す．第 4.6 章は結びである

4.2　日本語文書のジャンル分類

文書を構成する文章テキストには，その文書ジャンルに含まれる典型的な構成要素，テキストにおける構成要素の一般的な出現順序，構成要素間の関係などに特徴があるとされる．ジャンルに応じた特徴を利用し，文書の属するジャンルを推定することにより，当該文書の重要箇所抽出や要約文生成，表現の書き換えなどに重要な手掛かりを与えることになる．たとえば，論文ならば目的，方法，結果，考察，結論などの構造が，手紙なら，頭語，季節の挨拶，安否の問いかけ，本文，結語などから構成される．また，”文体” (style) とは，それぞれの作者がジャンルに応じて特有な文章表現上の特色を有し，文章の語句・語法・修辞などが作者固有の特徴・傾向となっているものをいう．

本研究では延べ語数での品詞分布を用いることで文書のジャンル類別が行えることを示す．文書の特徴としては語の分布が挙げられ，ジャンル分類では通常語の頻度が用いられる．小説や新聞などの系列長の長い文書では語彙数が数万から数十万となり，語の頻度による分類では数万次元の単語分布の比較が必要となる．これによりデータ数の増加に伴い必要なメモリ量が膨大となる．また，単語を特徴量として用いると語の共起による分類では多義語などの語の意味の扱いや，語の確率による分類では出現しない語の扱いにより分類結果が異なるという問題がある．

しかし，特徴量として品詞分布を用いることで，ジャンル分類に必要な次元数が語彙数から品詞数へと大幅に削減でき語の扱いが容易になる．また，日本語文書の品詞分布には法則性があることが知られており [56][54]，各ジャンルの品詞の割合の理論値は品詞分布の決定要素である名詞の割合によって求めることが可能である．文章を構成する主語・述語には，修飾語などの要素が付け加わって，より複雑な構造を形成する．文章を成り立たせる要素を「文の成分」というが，通常「主語」「述語」「修飾語」(連用修飾語・連体修飾語)「接続語」「独立語」の５つが挙げられる．形態素的には，名詞や動詞など独立した意味を担う自立語と，格などでこれらを補完する補助語があり，個々の語の利用を調べるだけではなくて，これらの分布を特徴量として用いることで，計量的な判断を高精度にすることが考えられる．

この法則性を用いることで品詞分布により各ジャンルの特徴を表すことができ，ジャンル分類が行えると考えられる．ジャンル分類では，各ジャンルに確率モデルを与え，品詞分布の生成過程を表現する．名詞の割合の事前分布にガウス分布を仮定し，ジャンルごとにガウス分布のパラメータを学習することで各ジャンルの名詞分布の特徴を得る．このガウス事前分布による各ジャンルへの所属確率と確率モデルから求まる理論値と観測値の誤差によりジャンル分類が行えることを示す．

4.3　日本語文書の品詞分布

日本語文書では，自立語と付属語が組み合わされて出現する．自立語とは独立して意味を成す語であり，典型的には名詞や動詞がある．一方付属語とは単独では意味を持

たず自立語と合わせて節を成す語であり，助詞や助動詞が挙げられる．例えば「出た出た月が」を単位語に区切ると「出/た/出/た/月/が」となり，自立語は「出」と「月」，延べ語数は 6 である.「出る」が 2 ,「た」が 2 ,「月」が 1 ,「が」が 1 であり，これはそれぞれの語の使用度数を示す．単位語の総数から，重複するものを省いて数えた語のが異なり語である．この例では，延べ語数は 6 であるが,「出る」と「た」が重なっているため，異なり語数は「出る」「た」「月」「が」の 4 つである．

樺島の法則は，延べ語数に関する品詞分布の特性である．自立語を名詞，動詞，名詞・動詞などについてそれがどのようであるかというありさまを示す形容詞類（形容詞・副詞・連体詞），文の成分としては比較的独立性をもつ接続詞類（接続詞・感動詞）に分ける．このとき，名詞の百分率を N，動詞を V，形容詞類を A_d，接続詞類を I としたとき各品詞の割合は以下の近似式で表される．

$$A_d = 45.67 - 0.60N \tag{4.1}$$

$$logI = 11.57 - 6.56logN \tag{4.2}$$

$$V = 100 - (N + A_d + I) \tag{4.3}$$

樺島の法則で示されている近似式は名詞の割合に対しての品詞分布の特性である．品詞の割合には文書のジャンルによって差がある [55] としながら，ジャンルによる品詞分布の特性を考慮していないため，当てはまりがあまり良くない．

文書ジャンルを考慮した品詞分布の法則としては，古典文学作品での品詞分布に関する大野の語彙法則がある．大野の語彙法則は異語数に関するものではあるが，古典 9 作品において名詞の割合がもっとも高い文書を左端に，名詞の割合がもっとも低い文書を右端におき二つの作品の品詞の割合を直線で結ぶと，他の作品の品詞の割合はこの直線状に垂直に並ぶ．ただ，大野の研究では古典文学作品のみに焦点を当てており，他のジャンルについてこの法則が成り立つかは論じられていない．

大野は語彙法則をグラフで図示したが，この定式化が水谷によってなされた [60]．ある 3 作品の名詞構成比を X_0，x，X_1 他の語類の構成比を Y_0，y，Y_1 としたとき以下の式が成り立つ．

$$y = Y_0 + \frac{Y_1 - Y_0}{X_1 - X_0}(x - X_0) \tag{4.4}$$

先行研究が着目するポイントは，延べ語数，異語数に関して日本語文書には品詞分布に法則性が存在することにある．本研究では品詞分布の特性がジャンルごとに変化すると仮定し，文書のジャンルごとに回帰直線を設ける．これにより品詞分布での分類が名詞により相関分析でき，名詞分布によりジャンル分類が行える．

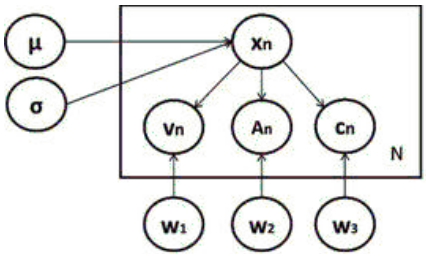

図 4.1: グラフィカルモデル

4.4 提案手法

4.4.1 品詞分布の確率モデル

本稿では，品詞分布の対応関係を確率モデルで表現し，ジャンル分類に用いる (図 4.1 で表す)．ここで X は名詞の割合，V は動詞の割合，A は形容詞類の割合，C は接続詞類の割合である．w は多項式係数ベクトルであり，μ と σ はガウス事前分布のパラメータである．文書の名詞の割合は事前分布であるガウス分布によって決まり，ガウス事前分布のパラメータがジャンルごとの名詞分布の特徴を表す．動詞，形容詞類，接続詞類の割合は名詞の割合と多項式係数ベクトルによって求まる．この確率モデルを各ジャンルに与え，ジャンルごとにガウス分布のパラメータと多項式係数ベクトルの学習を行う．ガウス分布のパラメータは学習データの平均と分散より (5) 式で求まる．各品詞の多項式係数ベクトル w_i は $\{w_{i0}, w_{i1}\}$ で表され，(6) 式・(7) 式より求まる．

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n \quad , \quad \sigma = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu)^2 \tag{4.5}$$

$$w_{i0} = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2} \tag{4.6}$$

$$w_{i1} = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k y_k \sum_{k=1}^{n} x_k}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2} \tag{4.7}$$

4.4.2 ジャンルの分類

ジャンル分類では，ジャンルごとに学習したパラメータより，動詞，形容詞類，接続詞類の品詞の割合の観測値と理論値の誤差を求め，ガウス分布を仮定した名詞の出現確率の逆数をかけることによりジャンル分類を行う．これにより単語の頻度を用いる分類では数万次元の単語分布を比較する必要があったのに対し，本手法では各ジャ

ンルの名詞の出現確率と多項式係数ベクトルにより分類を行うため高速に分類が行える．各ジャンルでの名詞の出現確率はガウス事前分布のパラメータである μ_j，σ_j により以下の式で求まる．

$$N(j) = \frac{1}{(2\pi\sigma_j)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma_j}(x-\mu_j)^2) \tag{4.8}$$

各ジャンルの動詞，形容詞類，接続詞類の理論値はテスト文書 $\{d_1, d_2, ..., d_n\}$ の名詞の割合 $\{x_1, x_2, ..., x_n\}$ と多項式係数ベクトルによって $w_{ji0} + w_{ji1} * x_n$ で求まる．ここで j はジャンル数であり，i は品詞数である．テスト文書の品詞の割合を動詞 $\{v_1, v_2, ..., v_n\}$, 形容詞類 $\{a_1, a_2, ..., a_n\}$, 接続詞類 $\{c_1, c_2, ..., c_n\}$ としたとき，理論値との誤差は以下の式より求まる．

$$\begin{aligned} V(j) = &(v_n - (w_{j11}x_n + w_{j10}))^2 + (a_n - (w_{j21}x_n + w_{j20}))^2 \\ &+ (c_n - (w_{j31}x_n + w_{j30}))^2 \end{aligned} \tag{4.9}$$

ジャンル分類では品詞分布の理論値と観測値の誤差と名詞の出現確率の逆数によって求まる値が最も低いジャンルを分類結果とする．以下の式に基づいて分類を行う．

$$Ans = arg \min_j V(j) * \frac{1}{N(j)} \tag{4.10}$$

4.5 実験

本章では2つの実験を行う．第1の実験ではF検定を行うことにより，回帰式が品詞分布の特性を表すものであることを示す．回帰式がジャンルに固有のものとなればコーパスに依存せずジャンル分類が行える．第2の実験ではジャンル分類を行う．ジャンル分類ではベースライン手法として単純ベイズを用いて精度と実行時間の比較を行う．品詞分布による分類が分類法によらず行えるか検証する．また，(形態素ではなくて)語自体の頻度による分類として，名詞のみを対象とする Latent Direchlet Allocation (LDA) を用いる．LDA による分類とは，ジャンルのトピックの確率分布とテスト文書のトピックの確率分布を比較し，KL 情報量が最も小さいジャンルを推定結果とするものである [34]．単純ベイズでは尤度が最も高いジャンルを推定結果とする．

4.5.1 実験準備

実験には4つのコーパスを用いる．各コーパスは，5人の著者の小説を20作品ずつ計100作品 (表 4.1)，朝日新聞の2007年1月度を30日分，日本語話し言葉コーパス

(Disc3) を収録順に先頭から 100 文書，NTCIR-3 より 98 年度公開特許公報全文データを 100 文書 (表 4.2) である．特許データは発明の詳細な説明のみを抽出して用いる．また，これらのうち小説・話し言葉・特許は 50 文書，新聞は 20 日分を学習データとして用いる．実験データには茶筅による形態素解析を行う．比較手法である LDA のパラメータはトピック数 150，ディリクレ分布の初期値を α=0.01，β=0.01 とし，ギブスサンプリング繰返しは 500 回行う．実行時間の測定に用いる計算機は CPU Intel Corei3 1.33GHz，メモリ 4GB である．

表 4.1: 小説リスト

著者	作品名
夏目 漱石	坊っちゃん，文芸と道徳，ケーベル先生，硝子戸の中 彼岸過迄，一夜，こころ，草枕，行人，幻影の盾 道草，門，手紙，野分，三四郎，変な音 趣味の遺伝，二百十日，虚子君へ，落第
芥川 龍之介	アグニの神，或敵打の話，桃太郎，疑惑，犬と笛 歯車，英雄の器，二つの手紙，一夕話，報恩記 影，羅生門，開化の良人，河童，三つの窓 おしの，地獄変，蜃気楼，運，彼
太宰 治	老ハイデルベルヒ，玩具，グッド・バイ，走れメロス 逆行，女生徒，佳日，火の鳥，鴎，犯人 喝采，故郷，狂言の神，黄金風景，ろまん燈籠 正義と微笑，斜陽，嘘，兄たち，女の決闘
島崎 藤村	秋草，嵐，朝飯，旧主人，食堂，岩石の間 三人の訪問者，並木，千曲川のスケッチ，船 伸び支度，芽生，分配，ある女の生涯，藁草履 家 (上)，刺繍，再婚について，北村透谷の短き一生，家（下）
森 鴎外	あそび，興津弥五右衛門の遺書，阿部一族，かのように，雁 カズイスチカ，魚玄機，最後の一句，細木香以，じいさんばあさん 堺事件，文づかい，余興，みちの記，心中 食堂，二人の友，沈黙の塔，高瀬舟，鶏

表 4.2: 特許リスト

公開番号
特開平１０－１，...，特開平１０－１００

4.5.2 評価方法

回帰直線の評価には F 検定を用い，分類の評価には f 尺度を用いる．F 検定では，回帰による要因効果と誤差による変動要因から求まる値が有意水準以上であれば，回帰

直線と観測値の間に有意差が有り，回帰直線が品詞分布の特性を表しているといえる．ここで回帰による要因効果についての平方和 S_F と平均平方 V_F を以下の式で定義する．

$$S_F = \sum_{i=1}^{n} (\bar{Y} - Y_i)^2 \quad , \quad V_F = \frac{S_F}{n-1} \tag{4.11}$$

次に誤差による変動要因についての平方和 S_C と平方平均 V_C を以下の式で定義する．

$$S_C = \sum_{i=1}^{m} (y_i - Y_i)^2 \quad , \quad V_C = \frac{S_C}{m-n} \tag{4.12}$$

この時,

$$F = \frac{V_F}{V_C} \tag{4.13}$$

となる．自由度は分母に対して (全標本数-群数) であり，分子に対して (群数-1) である．

次に分類の評価方法としてf尺度を用いる．f尺度は再現率と適合率の調和平均であり，実際に正であるもののうち，正であると予測されたものの割合である再現率を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \tag{4.14}$$

また，正と予測したデータのうち，実際に正であるものの割合である適合率を次のように定義する．

$$P_i = \frac{a_i}{a_i + b_i} \tag{4.15}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この2つの式の調和平均であるf尺度を次のように定義する

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad f = Average_i f_i$$

4.5.3 実験結果

まず，回帰式の係数とガウス分布のパラメータを表4.3，4.4に示す．各パラメータはジャンルごとに異なっており，ジャンルの特徴を表している．F検定の結果を表4.5，4.6に示す．すべての値が有意水準5％でのF分布の基準値より大きい値となっていることから，有意水準5％で回帰直線と観測値には有意差が有り，回帰直線が品詞分布の特性を表していると言える．

次にジャンル分類の結果と実行時間を表4.7，4.8に示す．ジャンル分類のf尺度は提案手法で0.945，単純ベイズで0.899，LDAで0.99となる．(単語分布を用いた) LDAによる分類と比して，品詞分類を用いた手法では精度が低いが, 提案手法では遜色がないほど高精度に分類が行える．提案手法による分類ではすべてのジャンルのf尺度が高い値を示す．一方，単純ベイズでは新聞のf尺度が低い．LDA分類と比較すると，単語分布では新聞と特許のf尺度が比較的低いが，品詞分布による分類では逆に新聞と

表 4.3: 回帰式の係数

	小説	話し言葉	特許	新聞
動詞				
w_{11}	-0.461	-0.039	-0.815	-0.690
w_{10}	0.557	0.304	0.804	0.718
形容詞類				
w_{21}	-0.453	-0.368	-0.127	-0.276
w_{20}	0.375	0.297	0.132	0.251
接続詞類				
w_{31}	-0.086	-0.593	-0.057	-0.033
w_{30}	0.069	0.399	0.064	0.031

表 4.4: 各ジャンルの名詞分布の平均と分散

	小説	話し言葉	特許	新聞
平均	0.585	0.440	0.779	0.799
分散	$1.84 * 10^{-3}$	$5.07 * 10^{-3}$	$1.63 * 10^{-3}$	$1.12 * 10^{-4}$

特許の f 尺度が他のジャンルより高いことに注目したい．実行時間は，提案手法で 5.05 秒，単純ベイズで 4.85 秒，LDA で 24576 秒となる．提案手法と単純ベイズではほとんど差がないが，LDA では品詞分布による分類の約 5000 倍となっている．

4.5.4 考察

ジャンル分類では提案手法による分類で f 値が 0.945 と高い精度になっている．単純ベイズによる分類では特許との名詞の割合の平均値が近い新聞で f 値が小さくなっているが，提案手法ではすべてのジャンルで高い精度となっている．このため，提案手法では名詞の割合の分布が近いジャンルが存在してもジャンル分類が行えると考えられる．また，単語分布による分類においても新聞と特許の f 尺度は他のジャンルより低くなったが，提案手法では逆に他のジャンルより高くなっている．表 4.9 よりテスト

表 4.5: F 検定結果 (自由度 1,48)

	動詞	形容詞類	接続詞類
小説	61.94	19.77	58.08
話し言葉	33.27	54.08	88.11
特許	364.68	85.50	45.49

表 4.6: F 検定結果 (自由度 1,8)

	動詞	形容詞類	接続詞類
新聞	85.40	8.38	12.96

表 4.7: ジャンル分類結果

	新聞	小説	話し言葉	特許	全体
(提案手法)					
再現率	1	1	0.840	0.960	0.950
適合率	0.909	0.847	1	1	0.939
f 値	0.952	0.917	0.913	0.980	0.945
(単純ベイズ)					
再現率	1	1	0.82	0.88	0.925
適合率	0.667	0.833	1	1	0.875
f 値	0.800	0.909	0.901	0.936	0.899
(LDA)					
再現率	0.960	1	1	1	0.990
適合率	1	1	1	0.962	0.990
f 値	0.980	1	1	0.980	0.990

文書の動詞の群内分散を見ると，F 検定の値が大きい特許と新聞では分散が 1.91E-04, 8.09E-06 となり非常に小さい値となっている．分散が最も大きい話し言葉では分散が 1.25E-03 となり，この値は分散の最も小さい新聞の 155 倍である．提案手法で f 尺度の高くなったジャンルは群内分散の小さいジャンルであり，回帰式の当てはまりがいいために精度が高くなったと考えられる．分類の実行時間は品詞分布での分類に比べて，繰り返し学習が必要な LDA での分類は 5000 倍近い値となっており，品詞分布での分類は非常に高速に行える．

表 4.8: 実行時間 (秒)

	提案手法	単純ベイズ	LDA
学習時間	4.92	4.74	24575
分類時間	0.125	0.109	0.68
合計	5.05	4.85	24576

表 4.9: 動詞の群内分散

	小説	話し言葉	特許	新聞
群内分散	$4.04 * 10^{-4}$	$1.25 * 10^{-3}$	$1.91 * 10^{-4}$	$8.09 * 10^{-6}$

4.6 結び

本論文では，品詞分布の決定要素である名詞分布によるジャンル分類法を提案した．また，品詞分布の特性は文書のジャンルごとに異なっており，ジャンルごとの回帰直線を用いることにより品詞分布の特性が得られることを検定により示した．

第5章　検索語の意味抽出

5.1　前書き

近年，インターネットから大量の電子文書を獲得するための手法が多数提案されている．この獲得した文書に対して，分類や異常値検出，要約，類似判定 [34] が行われている．しかし，インターネット中の文書は急激に増加することから，それらを体系的に組織化することや一貫した方法で効率的に抽出することは困難である．

一般的に情報検索では，検索語が与えられたとき検索語に関連した文書を抽出する．TF*IDF による重み付けを用いることで単語の重みに関連した文書を抽出できる．しかし，検索語に関係ない文書が抽出されることがある．例えば，検索者が”アップル”，”パイ”，”チャート”という検索語により，”アップルパイ”の種類の”チャート”を検索を行うとする，しかし，TF*IDF ベースの検索では，”アップル”コンピューターの”パイチャート (円グラフ)”と誤解し，iphone の情報が書いてある文書を抽出する可能性がある．このため検索者の欲しい情報を抽出するには，”アップルパイ”と”アップルコンピュータ”を別々に特徴付ける必要がある．このため，本研究では以下の 2 つの問題に取り組む．

- 如何に検索語間の依存を捉えるか.
- 如何に検索の文脈を特定するか

文書の特徴抽出に関しては過去に多数の研究がある．文脈解析の観点から，Kurohashi らは検索語の構文と依存関係の解析手法を提案している [19, 33]．しかし，この手法は状況を限定している．単語のクラスタリングに関しては，Latent Semantic Indexing (LSI)[9] や probabilistic LSI (pLSI)[13] など多数の研究が行われている．これらの研究では，潜在変数を用いて文書をトピックの混合で表現している．また，確率論に基づく HMM[40] やトピックモデル [5] が提案されている．HMM は文脈などの文書の構造の特性を捉えることができる．トピックモデルでは，各文書はトピックの混合で表され，各トピックは単語の確率分布によって特徴付けられる．トピックは意味が明示されない単語のクラスタであるが，潜在的な意味に対応すると仮定する．Latent Dirichlet Allocation(LDA) は，トピックモデルの一種として多数用いられる．

情報検索に対してディリクレスムージンクを用いた LDA に基づく尤度手法が提案されている [45]．この手法では，文書と検索の関係を捉えているが，検索語間の依存関

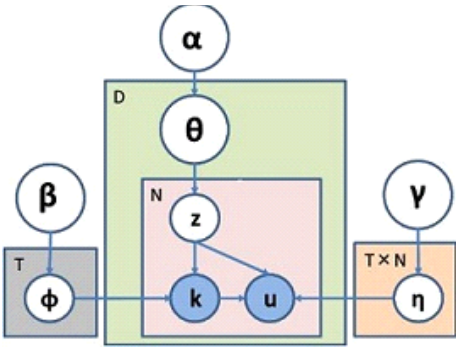

図 5.1: Graphical Model of Dependency Structure

係を考慮していない．検索者は検索の意図に基づき検索語に依存関係を与える．しかし，この単語のみの考慮では興味の内容を正確に捉えることは困難である．

本研究では，依存構造を用いて文脈を加味した検索手法を提案する．本研究の貢献は，主に以下の 2 点である．

- トピックモデルを用いた文書の文脈に重点を置いた新たな検索手法について提案する．
- 文脈上の依存構造を用いてトピックモデルの拡張を行う．

本稿は，以下の通りに構成する．第 5.2 章では，LDA の枠組みについて述べる．第 5.3 章では，文の依存構造について述べる．第 5.4 章では，依存構造によるモデル化について述べる．第 5.5 章では，実験によりモデルの有効性を示す．第 5.6 章では，関連研究について述べる．第 5.7 章で結びとする．

5.2 LDA の枠組み

本章では，LDA の枠組みについて述べる．単語 w は語彙 $\{1,\cdots,V\}$ の項目である．文書 d は，$d = \{w_1 ... w_N\}$ で表される N 個の単語の列である．ここで w_n は，文書内の n 番目の単語である．コーパスは D 個の文書の集合であり，$\{d_1, ..., d_D\}$ で表される．トピックモデルは，各文書が複数のトピックの混合として表現されるという仮定である．トピックは単語に関する多項分布を持つ．

言語モデル [3] は統計的手法として情報検索に用いられる．このモデルの枠組みの中で，クラスタベースのトピックモデルが提案されている [21]．混合ユニグラムモデルは，1 つの文書を 1 つのトピックから生成する．Probabilistic LSI (pLSI) は，潜在変数モデルとして多数用いられている．ここでは，文書・トピック・単語を確率的に対応付ける．Hoffman は，ベクトル空間モデルの枠組みの中で pLSI を検索に適用している．

pLSI では, LDA と違って, トピックの混合比を学習データの文書集合に依存して固定化している．一方, LDA ではこの混合比を事前分布から生成する点で異なる．LDA

のパラメータ推定では，潜在変数であるトピック z，文書ごとのトピックの確率分布 θ, トピックの単語分布 ϕ を文書集合に対して尤度が最大となるように推定する．

2 章図 2.1(a) では，LDA のグラフィカルモデルを示す. 図中の変数は, 図 2.1(a) 左下に, ディリクレ事前分布 $Dir(\beta)$，図 2.1(a) 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1(a) 左上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1(a) 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N_d は各文書の単語数を表す．LDA の単語生成過程を以下で示す. まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する.

Wei らは，LDA をスムージングに用いる検索手法を提案している．

$$P(w|d) = \lambda(\frac{N_d}{N_d+\mu}P_{ML}(w|d) + \frac{\mu}{N_d+\mu}P_{ML}(w|D)) + (1-\lambda)P_{lda}(w|d)$$

$$P_{lda}(w|d) = \sum_t P(w|t)P(t|d)$$

ここで N_d は文書 d の単語数である．μ はスムージングパラメータである．この式より，文書 d 内と全文書 D 内での単語 w の最尤推定量を求める．潜在変数の事後分布 $p(\theta, z_d|w) = p(\theta, z_d, w)/p(w)$ はギブスサンプリング [5] を用いて推定する. この手法では，トピックを用いて文書と単語の関係を捉えるているが，検索語間の依存関係を捉えることができない.

5.3 依存構造

本章では，フレーズの依存構造による文脈のモデル化とトピックモデルを用いた構造による文書の抽出手法について述べる．LDA を適用することで依存構造を抽出し，文脈を考慮した情報検索を行う．

5.3.1 文脈と依存構造

まず，文脈と依存構造の意味とモデル化について述べる．全ての文書は単語からなり，各単語は意味を持つ．例えば，”APPLE”という単語は，ある文書で”アメリカの多国籍コンピュータ企業”という意味を持つ．また，他の文書では，”バラ科リンゴ属落葉高木樹になる果実”を意味する．一般的に，単語は出現する文書に依存して複数の意味を持つ．各文書は複数の話題を含み，単語の分布はこの話題の影響を受ける．

本研究では，この単語の意味を捉えるために，係り受けを用いる．係り受けとは, 係り語と受け語の二項関係である．文節は，意味をなす最小の単位である． 文における任意の文節は, その文節の後の少なくとも 1 つの文節と係り受け関係を持つ [22]. 構文解析後, 文節では名詞句など一般的意味を成す語は内容語, 文節内で助詞など文法的な

役割を果たす語は機能語と呼ばれる．本研究では，検索語に名詞を用いるため，助詞以外を内容語として用いる．

例えば，(日本　新聞)，(日本　復活) という 2 つの係り受け語があるとする．前者の構造は新聞社を意味し，後者は経済関連を意味する．このように係り語”日本”の意味するところは，受け語”新聞”，”復活”によって変化する． この係り受けを用いることで検索の意味を特定することが可能になる．しかし，検索語に対する係り受け関係は一意に求めることができないという問題がある．

例えば，以下の文があったとする．

美しい 水車小屋 の 娘

この文は単語の繋がりによって 2 つの解釈ができる．

(美しい　水車小屋) の　娘
美しい (水車小屋　の　娘)

第 1 の構造は，”(美しい水車小屋)　の娘”を意味する．第 2 の構造は，”水車小屋の (美しい娘)”を意味する．ここでは，どちらの解釈も間違っておらず，文の曖昧性によって複数の解釈が存在する．このため，本研究では検索語間の依存構造を捉えることで検索の意味を特定する．提案手法では，トピックモデルにより，係り語を修飾する受け語を潜在的な状態として扱う．これにより, トピックと係り語の組から適切な受け語を選択する．

5.3.2　依存構造の推定

文書中の依存構造を扱うためには，文書中の構造の分布を解析する必要がある．しかし，文書中には様々な構造が混在しているため，特定の構造の情報を抽出することは困難である．本研究では，この構造をモデル化するために LDA に基づくモデル化手法を提案する．提案手法は，LDA に係り受け関係を組み込む．LDA はトピックモデルの一種であり，文書の生成過程を文書とトピックの依存とトピックと単語の依存関係により表現する．

提案手法のグラフィカルモデルを図 5.1 に示す．ここでは，係り受けの二項関係を用いて係り語 k とその受け語 u の組に対してトピックを推定する．LDA との違いは，提案モデルは単語の代わりに係り受け語を使用することである．また，この係り受け関係を表す新たな変数 η と γ を用いる．η はトピックと係り語ごとの受け語の偏りを表す多項分布である．γ は，η のディリクレ事前分布のパラメータである．係り受け語 k, u が事前分布から生成される確率は以下の式より求まる．

$$P(k, u|\alpha, \beta, \gamma) =$$

$$\int\int\int \prod_{z=1}^{T\times N} P(\eta_{z,n}|\gamma) \prod_{z=1}^{T} P(\phi_z|\beta)(P(\theta_d|\alpha)$$
$$(\prod_{n=1}^{N_d} \sum_{z_n}^{T} (P(z_n|\theta)P(k_n|z_n,\phi)P(u_n|z_n,k_n,\eta))))d\theta d\phi d\eta$$

潜在変数 z はギブスサンプリングにより推定する．これにより，文書ごとのトピック分布 θ，トピックごとの単語分布 ϕ，トピックと係り語の組が持つ受け語の分布 η をそれぞれ推定する．

$$P(z_i = j|\backslash i, \mathbf{k}, \mathbf{u}) = \frac{(n^{(u_i)}_{\backslash i,k_i,j} + \gamma)(n^{(k_i)}_{\backslash i,j} + \beta)(n^{(d)}_{\backslash i,j} + \alpha)}{(n^{(\bullet)}_{\backslash i,k_i,j} + \gamma)(n^{(\bullet)}_{\backslash i,j} + \beta)}$$

ここで，n^d_j は，文書 d にトピック j が出現した数, $n^{(k_i)}_j$ は，j に係り語 k_i が出現した数, $n^{(u_i)}_{k_i,j}$ は，トピック j の係り語 k_i に受け語 u_i が出現した数，$n^{(\bullet)}_j$ は，j が出現した総数, $n^{(\bullet)}_{k_i,j}$ は，j に k_i が出現した総数である．この提案モデルにより，文書集合をモデル化することで係り受け関係を学習する．

5.4 依存構造を用いた検索

本章では，依存構造を用いた意味を加味した検索手法について述べる．検索語は使用者によって与えられる単語の集合であり，各検索語は何らかの意味を持つ．提案手法は，検索語間に依存構造を仮定する．この依存構造により検索語の意味を推定する．例えば，”アメリカ”，”日本”，”ミーティング”という検索語があるとする．この各検索語に対して係り受け語を与えることで，”アメリカ　大統領”，”日本　総理大臣”，”サミット　ミーティング”という拡張された検索語を得る．この拡張した検索語に対して提案モデルにより，あるトピックでの出現確率を与える．全ての検索語に対して各トピックでの出現確率を計算し，単純ベイズでの結合確率を求める．最尤推定により最も確率の高い共通のトピックでの係り受け語を検索語として用いる．各文書は異なるトピック分布を持つため，検索語となる係り受け関係は文書ごとに異なる．

例えば，文書 d_1 と d_2 では，文書のトピック分布によって以下のように各検索語に対して異なる係り受け語が抽出される．

d_1 ： (アメリカ クラブ)，(日本 チーム)，(ミーティング 90 分)
d_2 ： (アメリカ 大統領)，(日本 総理大臣)，(ミーティング 見通し)

文書 d_1 で抽出された 3 つの係り受け語”(アメリカ クラブ)”，”(日本 チーム)”，”(ミーティング 90 分)”がトピックから出力される結合確率を検索語に対する推定値とする．同様に d_2 では，”(アメリカ 大統領)”，”(日本 総理大臣)”，”(ミーティング 見通し)”がトピックから出力される結合確率を推定値とする．この推定値を比較することにより，各文書をランク付けする．

5.5 実験

5.5.1 前処理と評価

実験データは，毎日新聞の 2009 年 1 月 1 日の先頭から 10000 記事と 20000 記事の 2 つのセットを用いる. 各記事は，日本語形態素解析ツールの JUMAN と日本語構文解析ツールの KNP[19] を用いて前処理を行う．実験データの詳細は，図 5.1 に示す．実験では，パープレキシティと類似度によりモデルの性能を評価する.

表 5.1: Mainichi News Corpus 2009

Item	Number	Per Article	Number	Per Article
	20,000	-	10,000	-
文字数	7,446,344	372.32	3,767,329	376.73
単語数	3,401,267	170.06	1,717,357	171.74
ライン数	137,579	6.88	67,825	6.78
Dependency	525,074	26.25	-	-
Occurrences	1,827,545	91.38	-	-

まず, モデルの評価を行い, トピック数と繰り返し回数を決定する. 言語モデルの性能評価に情報理論に基づく客観的評価手法であるテストセット・パープレキシティを用いる. テストセット・パープレキシティは以下の式で計算する.

$$PP = 2^{H(p)}, H(p) = -\frac{\sum_X \log_2 p(X)}{N}$$

ここで $X_1 \cdots X_N$ は評価用のテキスト集合とする．各語の確率を求め，そのすべての情報量の平均を EXP の係数としている．すべての語の生起確率を求めるには，トピックの確率×語の生起確率を用いる．P_M は言語モデル M による $X_1 \cdots X_N$ の生成確率を表す. パープレキシティ値が高いほど単語の特定が難しく, 言語として複雑である. よって, パープレキシティの値が低いほど言語モデルの性能が高いと評価できる.

$$P_M(X_1 \cdots X_N) = \Pi_{i=1}^{N} p(X_i)$$

$$PP(s) = P_M(X_1 \cdots X_N)^{-\frac{1}{N}} = 1/\sqrt[N]{p(X_1) \cdots p(X_N)}$$

次に，トピック内での文書間の類似度を測る．文書の単語に TF*IDF による重み付けを行い，余弦類似度を測る．ここでは，閾値を設けて関連が疎な文書を除外する.

最後に新聞記事に対して検索を行う．ここでは，20000 記事の中から 10 記事を正解記事とする．この正解記事の見出しから検索語を抽出する．検索語の数が 2 の場合と 3 の場合でそれぞれ精度を比較する．表 5.2 は，用いる検索語と見出しである．この検索語を用いて 20000 記事をランク付けする．比較手法には，標準的な TF*IDF を用いる [11].

表 5.2: 検索対象となる 10 記事

No.	見出し	検索語
1	大分コンサル脱税:暴力団対策に 5 億円 大賀容疑者、腹心に報酬	大賀, 容疑, 脱税
2	クリントン国務長官:麻生首相と会談、日米同盟強化で一致　小沢・民主代表とも	会談, クリントン, 長官
3	東京・江東の女性バラバラ殺害：星島被告、無期懲役「死刑重すぎる」－－地裁判決	無期, 死刑, 懲役
4	日露首脳会談：領土問題「政治が決断」　首相、新アプローチ【大阪】	首脳, 会談, 露
5	09 年度予算案：衆院審議空転、月内通過も黄信号　いら立ち募らせる与党	予算, 審議, 衆院
6	ゼロゼロ物件：未明督促に賠償命令　家賃保証会社に－－福岡簡裁	保証, 家賃, ゼロゼロ
7	体外受精：受精卵取り違え？妊娠　２０代女性中絶－－香川の県立病院	卵, 受精, 病院
8	あしたのジョー：ボクシング漫画、「週刊現代」に復刻連載	復刻, 連載, あした
9	受精卵取り違え：夫婦「１００％我が子なら…」　調査法なく出産断念	卵, 受精, 夫婦
10	かんぽの宿：「２年以内の譲渡可能」契約ただし書き指摘－－総務相	宿, 譲渡, かんぽ

5.5.2　実験結果

まず，第 1 の実験より，10000 記事と 20000 記事でのパープレキシティを表 5.3，5.4 に示す．T はトピック数である．

表 5.3: テストセットパープレキシティ (10000)

	Iteration				
T	200	400	600	800	1000
50	6466.88	6184.02	6110.86	6070.93	6054.05
100	5048.64	4873.80	4791.86	4744.52	4721.45
150	4501.88	4333.43	4281.81	4257.28	4238.26
200	4181.79	4073.29	4033.04	4014.49	3993.93

表 5.4: テストセットパープレキシティ (20000)

	Iteration				
T	200	400	600	800	1000
50	7962.55	7639.58	7509.45	7449.90	7418.68
100	6229.89	6026.60	5963.95	5932.45	5905.87
150	5556.85	5373.78	5309.00	5277.59	5260.03
200	5329.33	5176.40	5138.58	5105.77	5071.59
250	5122.78	4989.31	4950.89	4934.12	4920.05
300	5035.64	4931.95	4889.32	4862.85	4856.43

表より，パープレキシティは 10000 記事と 20000 記事のどちらも繰り返し回数が 400 を超えた付近から変化が減少している．繰り返し回数 1000 では，どちらもほとんど収束している．この結果から，10000 記事では繰り返し回数 1000，トピック数 200 とし，20000 記事では，繰り返し回数 1000，トピック数 300 とする．

表 5.5 は，10000 記事での各トピックの類似度である．図 5.2 は，各トピックの文書数である．図 5.3 は，各トピックのコサイン類似度である．

全体の類似度の平均は 0.0217 であり，各トピックの平均は 0.0315 である．最大の類似度はトピック 15 の 0.1923 である．最小の類似度はトピック 158 の 0.0174 である．200 トピック中 144 のトピックが全体の類似度の平均を上回っている．

表 5.6 は，提案手法より各トピックで上位 10 組となった係り受け関係である．表より，トピック 1 は，主に列車に関する単語を含んでいる．トピック 2 は，就職関係の単語を含んでいる．トピック 3 は，芸術や劇場に関する単語を含んでいる．

表 5.7 には文書検索の結果の順位を示す．表より，提案手法は TF*IDF と比較して，検索語の数 2 の場合で 5 記事に対して結果が改善されている．検索語の数が 3 の場合で 6 記事に対して結果が改善されている．どちらの場合も 1 記事は同等である．

表 5.5: 各トピック内での類似度

Topics	記事数	類似度
200	53,388	0.0217
per topic	266.9	0.0315
maximum	3246	(0.0252)
	(2284)	0.1923
minimum	74	(0.0433)
	(144)	0.0174

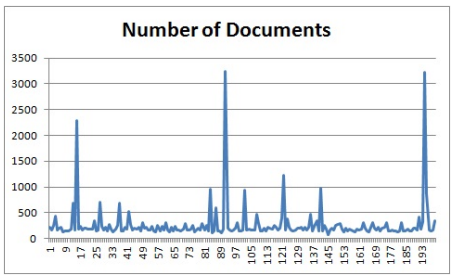

図 5.2: 各トピックの文書数

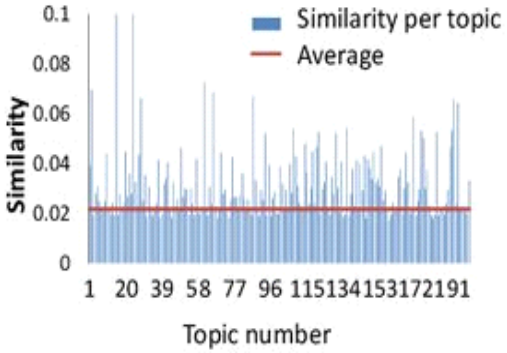

図 5.3: 各トピックのコサイン類似度

5.5.3 考察

まず，第 1 の実験結果から，提案モデルは，繰り返し回数 400 を超えた付近から収束し始める．次に，提案手法のトピック内の類似度は全トピックの 72%が平均余弦類似度を上回っている．このことから，提案モデルによる文書集合のモデル化が適切に働いていると考えられる．

表 5.8 には 200 トピック中で最も類似度が高くなったトピックと低くなったトピックの係り受け語の詳細を示す．

最も類似度が高くなったトピックでは，国内政治に関する単語が多く出現している．さらにこのトピックを含む文書同士は類似度が高くなっている．一方，類似度が最も低くなったトピックでは，全体としてまとまった意味を有していない．また，このトピックを含む文書間の類似度は低くなっている．

表 5.9 は，各トピックでの係り語”日本”を含む係り受け関係である．この結果より，係り語に対するトピックは受け語の文脈によって変化している．トピック 1 の受け語は，航空会社関連の語が集まっている．トピック 2 の受け語は，国際政治関連の語が集まっている．また，表 5.10 は，各トピックで確率の最も高くなった係り受け関係である．トピック全体としても，トピック 1 は航空会社関連の係り受け語が出現してい

表 5.6: 依存構造

Topic 1		Topic 2		Topic 3	
係り語	受け語	係り語	受け語	係り語	受け語
列車	旅	派遣	切り	愛	劇場
影響	出た	失職した	農民工	根強い	人気
全貌	見合わせ	正社員	して	史緒	萌
見合わせ	影響した	雇用	守る	昭和	におい
売却	検討対象	雇う	いう	男	女
運転	見合わせ	労使	合意	兄	帰る
提案コンペ	実施	短縮する	こと	あり	ふれた
証券主要	5 社	事業	主に	俳優	河原雅彦さん
多額の	損失	向き合い	展開すべきだ	透き通る	素材
反対側	ドア	派遣期間	上限	上演	なる

表 5.7: 単語数 2，単語数 3 でのランク付け; ”*” は固有名詞を含む

No	2 Words		3 Words	
	TF*IDF	LDA	TF*IDF	LDA
1*	9	6	7	5
2*	2	6	2	5
3	2	4	2	1
4	6	3	6	1
5	12	3	12	1
6*	1	8	1	7
7	3	1	3	1
8*	1	15	1	10
9	1	1	1	1
10*	16	4	16	2

る．同様に，トピック 2 は，国政政治関連の係り受け語が出現している．

文書検索の結果は，検索語の数 2 の場合では，TF*IDF と比較して改善 5，悪化 4 となり，検索精度が若干改善している．検索語の数 3 の場合では，改善 6，悪化 3 となり，提案手法のほうが 2 倍改善している．

検索番号 4 の見出しは，”日露首脳会談 : 領土問題「政治が決断」　首相、新アプローチ【大阪】”である．検索語は TF*IDF 値により”会談”，”首脳”，”露”の 3 つとなる．各検索語を含む係り受けは表 5.11 に示す．この検索語の係り受けのトピック内上位 5 個の係り受けを表 5.12 に示す．表 5.12 は，北方領土問題に関連するトピックである．見出しとの比較から検索語の係り受けと意味が一致している．検索番号 4 に対するベクトル空間モデルでの検索において，”会談”，”首脳”，”露”の TF*IDF 値はそれぞれ 37.495，30.3249，28.051 となっている．ベクトル空間モデルのランク付け 1 位の文書の見出しは「日米首脳会談 :「親密さ発信」思惑空振り　昼食会も共同会見もなし」となっている．この見出しには”露”が含まれない．これは”会談”の TF*IDF 値が他の 2 つと比較して高い値であるため，この検索語の影響により誤ったランク付けが行われたと考えられる．提案手法はトピックと係り受けを考慮することで，適切にランク付けが行えている．

検索語と文書検索の結果から，TF*IDF は検索語に固有名詞が含まれる場合に高精

表 5.8: 平均余弦類似度の最大・最小類似度の係り受け

最大類似度		最小類似度	
係り語	受け語	係り語	受け語
突入する	表明	仮眠中	流された
全面対決姿勢	突入する	脱官僚	地域主権
対韓国窓口である	祖国平和統一委員会	斎藤さん	父
南北首脳宣言	07 年	兄弟子	3 人
動き	ある	同保安倍	よる
肯定	否定	第 8	き
核保有国	して	かおる	被告
打ち上げる	こと	肖像画	えり
ミサイル	怖い	親分	顔
無効	宣言した	相撲	続ける

表 5.9: 「日本」に関する係り受け

Topic 1		Topic 2	
係り語	受け語	係り語	受け語
日本	航空会社	日本	首相
日本	考えられる	日本	訪問団
日本	出発する	日本	支援団
日本	空	日本	とって

度に推定可能である．これに対し，提案手法は，検索語に固有名詞が含まれていない場合でも正確な推定ができている．

5.6 関連研究

近年多数の先行研究で情報検索においてトピックモデルが用いられている．Wei らは，アドホック検索に対して，LDA に基づく文書モデルを提案している [45]．ここでは，検索語の尤度をディリクレスムージンクによる文書モデル [53] と LDA の結合によって与える．Yi らは，検索問題に LDA や PAM[20] といった異なるトピックモデルを様々な検索問題に適用したときの性能ついて調査している．これらの先行研究は既存のトピックモデルを検索に使用しており，検索を目的としたモデルを構築していな

表 5.10: 2 トピックにおける上位の係り受け

Topic 1		Topic 2	
係り語	受け語	係り語	受け語
日本	航空会社	北方	四島
飛行計画	承認	出入国カード	提出
離陸	始めた	戦後	首相
完全に	追い払うの	日本	首相
けが人	無かった	帰属問題	最終的解決

表 5.11: 係り受け (検索番号 4)

係り語	受け語
個別会談	行うつもりだ
日露首脳会談領土問題	大阪
日露双方	関心事項

表 5.12: 係り受けのトピック内での上位 5 個

係り語	受け語
北方	四島
出入国カード	提出
戦後	首相
日本	首相
帰属問題	最終的解決

い．単語間の依存を捉えることができるが，検索語間の依存を考慮できない．このため，意味を考慮した情報検索に対するモデル化としては不十分である．

トピックモデルの問題点の一つは関連語や高頻度語の共起関係のモデル化の欠落にある．実際，どの文書にも出現する一般語といった高頻度語はモデルの性能を悪化させる要因になる．ディリクレスムージングは文書と検索語の関係を捉えるのに有用であるが，検索語間の依存性を捉えるには不向きである．

先行研究では，LDA に対するオンライン学習手法 [7, 14, 32] や時系列トピックモデル [6, 39] が示されている．Blei らは，LDA に対する変分ベイズ法を用いたオンライン学習手法を示している．本研究の提案手法においても，オンライン学習手法を用いることでデータストリームといった動的な文書集合に適用することが可能となる．

5.7 結び

本研究では，依存構造を用いて文脈を加味した検索手法を提案した．ここでは，係り受け関係により LDA を拡張することで検索語間の意味を推定した．

実験結果より，各トピックには似た文書が多く含まれており，検索を行うことが困難であることを示した．ベクトル空間モデルは固有名詞が検索語に含まれる場合に高精度となる．提案手法は，トピックモデルにより文脈を捉えることでより高精度になることを示した．提案モデルでは全トピックの 72%が全体の平均余弦類似度を上回ることを示した．これにより，提案モデルでは文書の特徴を捉えられることを示した．

第6章　事前分布の学習による動的な特徴抽出

6.1　前書き

近年，インターネットの発達から大量の文書を入手することが容易になっている．また，ブログ，ツイッター，LINEやフェイスブックに代表されるビッグデータやソーシャルネットワークサービスには多種多様な情報が含まれているが，未整理の大量の文書を人手で分類することは困難である．このため自動的に分類を行う手法が必要である．特に，ニュース記事，マイクロブログに代表されるストリームデータは極めて重要な情報源として注目を集めている．これらの文書集合ではリアルタイムな情報を扱うことから，ストリームデータの解析を行うことで現在の流行や評判情報など様々な知識を抽出したい．

次々に新たな文書が到着するという文書ストリームのような動的な文書集合では，話題の変化に応じて分類先となるクラスの特徴が変化することから，分類基準を固定することは困難となる．このようなストリームデータは新しいデータが逐次到着するためデータ量が膨大となり，集合で扱われる話題が動的に変化する．例えば，政治に関する話題では，選挙，予算編成，法案審議のような時期に依存したイベントや，外交問題や汚職事件など不定期なイベントに依存し，発生する単語分布は大きく異なる．これらのイベントの中でも，時間経過によって論じられる内容は逐次変化している．このため，文書ストリームの分類では，準教師有り学習のように文書集合の変化を新たに到着した文書から学習し，分類基準を動的に変化させる適応的な分類手法が必要となる．

文書をクラスごとに分類する文書分類では，政治，スポーツ，ITといったクラスは文書の集合として表現され，文書の様々な特徴によって同一のクラスとして同定される．一般的にトピックモデルは，単語の統計に基づいて直感的にクラスの特徴を捉えるため様々な用途に用いる．一方で，文書の話題は独自の情報や特徴を持ち，固有な単語やそれらの分布によって話題独自の特徴を有する．文書はクラス内の特定の話題について書かれているとき，話題の影響によって特定の単語が多く出現する．文書は一般的に異なる特徴を持つ複数の話題によって構成される．

ニュースストリームでは，特定の話題が急激に生じるバースト現象が生じる[18]．例えば，衆議院選挙のような大きなイベントが起これば，選挙期間中は他の話題に関する記事が減少する代わりに，衆議院選挙に関する記事が急増し，選挙が終わると異な

る話題に遷移していき選挙に関する記事は減少する．この話題の変化によってクラスの特徴が変化していくため，動的な学習により分類基準を変更する必要がある．

本研究では，文書ストリーム中の文書を動的にカテゴリ分類するため，2 つの問題を論じる．第 1 の問題は，特徴の変化に伴って如何に適切な特徴を学習するか，第 2 の問題は，特徴の変化が起きる文書ストリームで各話題に対する適切な特徴を学習するかである．トピックモデルは確率論の観点から文書集合のモデル化を行う手法である．このモデルでは文書をトピックの混合として，各トピックを単語についての確率分布として表現する．これらの分布はディリクレ事前分布によって制御する．近年トピックモデルを用いた多くの研究が行われ，著者推定 [34] や文脈を考慮した情報検索 [36] 等の応用例がある．特にオンライントピックモデルは動的学習により，話題の変化に応じて新たな特徴を捉えることができる．また，Ramage らは学習文書に対するトピックモデルの適用として，トピックとクラスを 1 対 1 で対応させることでモデル化を行う手法が示されている [28]．McCallum ら [23] は，ガウス分布を前提とした EM アルゴリズムを用いて，文書分類で多重クラス推定を行う方式を提案している．ここでクラスラベルの組合わせ (文脈) の個々に対して，混合分布 (各確率分布パラメタと混合比) を計算し，最大尤度のクラスラベルを割り当てる．しかし，文脈はクラス数に対して指数個あるため，計算時間が膨大となる．

本研究では，オンライントピックモデルと部分学習文書ストリームを用いて，ニュースストリーム中で動的に文書分類を行う方式を提案する．部分学習文書ストリームとは，ストリーム中に生じるニュースの一部がクラスラベルを有するものをいう．提案する手法では，話題のバーストに対処するために，話題の発生確率の事前分布を設ける．オンライントピックモデルを各クラスに対して，独立して適用することで単語分布の特徴を抽出する．さらに，部分学習ニュースストリームでは，話題の変化に応じて各パラメータを推定することにより，分類精度を保ったままで分類を行う．本論文の貢献は主として 2 つの点で重要である．即ち，(1) オンライントピックモデルと部分学習文書を用いてバースト特性に重点を置いた新たな分類手法を提案していること，および (2) 定常確率分布に対して適切なパラメータを求めるだけではなく，求める真の分布が変化する確率分布に対して動的に適応させることである．

第 6.2 章では文書ストリームについて述べ，第 6.3 章ではトピックモデルについて述べる．第 6.4 章では提案手法について述べ．第 6.5 章では実験により有効性を示す．第 6.6 章で結びとする．

6.2 文書ストリーム

ニュースストリーム，マイクロブログに代表される SNS データは，極めて重要な情報源として注目されており，娯楽，広告や企業分析等の目的で多数の人々が利用している．これらのデータは日々増加しており，話題の中心は逐次変化する．このような文書ストリームでは，文書は無限に増加するため，文書をカテゴリ分類する必要性は

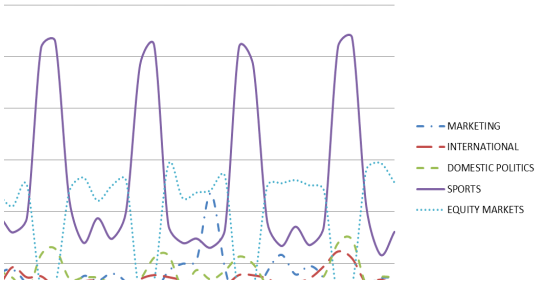

図 6.1: Frequency of class

高いが，反面過去の文書全てを用いた一括学習は，性能からも話題の即時性を捉えにくい点からも対応が困難である．このため，文書ストリームを扱うには，出現した文書を一度だけ学習に用いる単一パスで学習を行える動的な手法が必要である．

ニューズストリームでは，一度新たな話題が発生すると関連した記事が短期間に多数発生し，単語分布が大きく変わるというバースト現象が発生する [18]．各話題に依存した単語分布は，新たなイベントの影響を強く受ける．これが話題のバーストを生じる原因である．例えば，図 6.1 より Reuters Corpus の 1996 年 8 月 20 日から 30 日間のクラスの出現確率では，全体としては SPORTS の出現の出現が多いが一定値を保っているのではなく，EQUITY MARKETS と交互に出現する．これらのクラスでは出現確率が周期的に変化している．MARKETING では出現確率が非周期的に変化しており，全体としては低い出現確率になっているが，最も出現確率の高いクラスと同程度の確率となっている時期もある．

ニューズストリームでは，各記事は記事が属するクラスに関するいくつかの話題を論じているが，この話題はストリーム中で大きく変化する可能性がある．即ち，クラス内であっても単語分布は各話題に関連して多様性を持つ．このような時間に応じて特徴が変化する集合に対して，変化に対応した適切な特徴を抽出するためには動的学習を行う必要がある．学習データ (クラスラベルの付いた文書) と新たに出現するテストデータを同時に使用することにより，適応的に特徴を抽出する．反面，準教師有り学習では，誤った特徴が学習され，分類の精度が悪化する可能性がある．しかし，新たに到着した文書全てを人手によりラベル付けすることは，コストが膨大となるため困難である．

本研究では，学習データとテストデータが混在する部分学習文書ストリームから如何に動的学習を行うかといった問題を扱う．新たに到着した全ての文書をラベル付けすることは困難であるが，一部の文書のみをラベル付けすることや少量の学習データ (クラスラベル付き文書) をストリーム中から獲得することは可能である．例えば，ツイッターのツイートストリームの分類でハッシュタグから得られるラベルをクラスとした場合，ストリーム中から新たな学習データを得る．また，同じ話題についてのツ

イートであっても，多くのツイート文書ではハッシュタグは付与されておらず，獲得できる新たな学習データはハッシュタグの付いた少量のツイートとなる．

部分学習文書ストリーム内の文書分類では，学習データとテストデータの双方を用いてパラメタ更新 (学習) する．クラスの単語分布 $V = v_1, v_2, ..., v_n$ は学習文書の単語分布 V^{train} とテスト文書の単語分布 V^{test} の線形結合で表す.

$$V = (1 - \lambda)V^{train} + \lambda V^{test} \tag{6.1}$$

λ は学習率であり，クラスの単語分布を構成するテスト文書と学習文書の比率を表す．学習データだけを用いた更新では，話題のバーストが生じたクラスの変化に対応できない．実際，文書ストリーム内の学習データは (テストデータと比して) 少量であるため，単語分布に対する影響が少ない.

6.3 トピックモデル

本研究ではトピックモデルに基づく部分学習文書による分類問題を論じる．トピックモデルは文書集合の分析に非常に有効なモデルとして注目されており，文書分類や文書要約など幅広く用いられる．トピックモデルとは, 1 つの文書は特定の話題 (内容) を表す．話題はいずれかのカテゴリに属し，カテゴリはラベル (あるいはクラス) で記述される．このとき話題は複数のトピックの混合として表現される．1 つの文書が 1 つのトピックで表される多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある. その中でも最近用いられているのが LDA(Latent Dirichlet Allocation) である．確率的潜在意味索引付け (Probabilistic Latent Semantic Indexing, pLSI) は, LDA と違って, トピックと単語の多項分布にそれぞれディリクレ事前分布を導入し, 混合比を学習データの文書集合に依存して固定化している [13]. 一方, LDA ではこの混合比は事前分布から動的に生成する点で異なる. LDA は学習データだけに依存しないが, 代わりに特定の確率モデルを仮定するため, 柔軟な観点で文書を扱える可能性がある.

トピックモデルをニュース記事集合に適用することで，クラスをトピックの分布で表し，クラス内の複数の話題を特徴付けることが可能となる．話題は記事の内容を意味し，トピックは特定の意味を持たない単語のクラスタであるが，話題はトピックに対応すると仮定する．これにより話題の変化をトピックの変化として表現する.

2 章図 2.1(a) は LDA のグラフィカルモデルである. 図中の変数は, 図 2.1(a) 左に, ディリクレ事前分布 $Dir(\beta)$，図 2.1(a) 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1(a) 上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1(a) 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N は各文書の単語数を表す. LDA の単語生成過程を以下で示す. まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する.

次に, 文書d内のi番目の単語w_iにおいて，抽出した文書dの多項分布$Multinomial(\theta_d)$からトピックz_iを抽出し，そのトピックz_iの多項分布$Multinomial(\phi_{z_i})$から単語w_iを抽出する.

6.4 オンライントピックモデルを用いた分類

6.4.1 動機

トピックモデルではパラメータ学習を一括して行う．しかし，大量の学習データに対しては極めて効率が悪い．この問題を解決するためにオンライントピックモデルが注目されている [7, 14, 32]．オンライントピックモデルでは，差分的なギブスサンプリングやEMアルゴリズムによるパラメータ学習により，一括学習の適用が困難な膨大な文書集合を効率よく扱う．また，ストリームデータなどの動的な文書集合に適用すれば，集合の変化に応じてパラメータを更新することで，新たな特徴を学習することができる．Yaoらによるオンライントピックモデルでは [50]，差分型ギブスサンプリングによりパラメータの更新を行い，文書分類を行えることを示している．しかし，ここではトピックモデルのパラメータを更新するのみであり，ストリームデータの特性を考慮していない.

本研究では，オンライントピックモデルに基づくニュースストリームに対する新たな分類モデルを提案する．ここでは文書が各ラベルを有する (クラスに属する) 確率分布 (クラス出現確率分布) に事前分布を仮定し，ストリーム中の各クラスに対して独立にオンライントピックモデルを適用するという混合オンライントピックモデルを使用する．本研究の貢献は，定常な確率分布に対してパラメータの調整を行うのではなく，文書集合の適切なパラメータを求め，提案手法の枠組みの中で各確率分布を動的に適応させることである.

6.4.2 オンラントピックモデルの構築

本研究では，各クラスについてのディリクレ事前分布と各クラスの文書の特徴を学習するために，オンライントピックモデルを適用する．オンライントピックモデルは当該クラスでの文書ストリームを扱い，クラス内での語・文書の分布変動を記述できる．バースト現象はクラスごとに生じ，変動が他クラスに影響を与えない．どのクラスでバーストが生じるかは多項分布確率的に生じると仮定する．従って複数のオンライントピックモデルを混合して協調する上位モデルが必要である．以下では混合オンライントピックモデルの構成を論じる．まずクラスの特徴を学習するために各クラスの文書集合にトピックモデルを適用する．ここで各クラスの文書集合は独立しており，各モデルのトピック分布，トピックごとの単語分布は，他のクラスからの影響を受けない．次に各クラスの学習文書に対してギブスサンプリングによりトピック分布と単

語分布に対する各パラメータを推定し，学習されたパラメータを用いて文書を分類する．続いて分類結果を基に学習を行うことで新しいパラメータを得る．この過程を繰り返すことで文書ストリーム内の文書分類を行う．ここで各パラメータは他のクラスとは独立に決定される．まず初めに，ギブスサンプリングの結果，推定されるトピックごとの単語分布 ϕ とクラスのトピック分布 θ の値は以下の式で求める．

$$P(z_i = j | w_i = m, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto \phi_{mj} \times \theta_{dj} \tag{6.2}$$

$$\phi_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta} \tag{6.3}$$

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha} \tag{6.4}$$

C_{mj}^{WT} は単語 m がトピック j に割り当てられた回数，C_{dj}^{DT} は文書 d がトピック j に割り当てられた回数，V は全単語数，T は全トピック数である．この学習されたパラメータを用いてテスト文書を生成する尤度を計算することで分類を行う．

続いて分類結果を基にテスト文書が分類された先のクラスのパラメータを更新する．パラメータの更新には差分型ギブスサンプリングを用いる．差分型ギブスサンプリングでは，現在までのパラメータを用いて新たに到着したテスト文書のサンプリングを行い，当該文書内のみで繰り返し学習を行う．また，到着した文書が学習文書であった場合，該当のクラスのパラメータを更新する．

ここでトピックごとの単語分布 ϕ とクラスのトピック分布 θ は学習文書の $\phi^{train}, \theta^{train}$ とテスト文書の $\phi^{test}, \theta^{train}$ の線形結合で表す．

$$\theta_c = (1 - \lambda_c)\theta_c^{train} + \lambda_c \theta_c^{test} \tag{6.5}$$

$$\phi_c = (1 - \lambda_c)\phi_c^{train} + \lambda_c \phi_c^{test} \tag{6.6}$$

クラスの出現確率は一定間隔ごとに行い，本研究では 1 日単位で更新を行うこととする．時刻 t でのクラスの出現確率は時刻 t-1 でのクラスの出現確率とディリクレ事前分布のパラメータである γ を用いて推定する．クラスの出現確率は以下の式より求める．

$$P(c_t) = \frac{n_c^{t-1} + \gamma_c}{n^{t-1} + \sum_c \gamma_c} \tag{6.7}$$

クラスの出現確率の事前分布である γ は，学習率を決定するためのパラメータでもあり，学習率 λ_c は以下の式によって求める．

$$\lambda_c = \frac{\gamma_c}{\sum_c \gamma_c} \tag{6.8}$$

6.4.3　分類とディリクレ分布の更新

文書ストリーム中の文書を分類するため，新たに到着した未知文書に対して，各クラスから生成される尤度を求める．分類には提案モデルによって学習した各クラスのパラメータを用い，最も尤度の高いクラスに分類する．ここでは記事分類に加え，適切なディリクレ分布の差分学習も同時に行う．尤度は以下の式より求める．

$$Ans(c) = arg \max_{c} P(c;\gamma_c)P(d|\phi_c,\theta_c,\alpha_c,\beta_c) \tag{6.9}$$

$P(c;\gamma_c)$ はパラメータ γ_c を基にしたクラスの出現確率，$P(D|\phi_c,\theta_c,\alpha_c,\beta_c)$ は各クラスから未知の文書が生成される尤度である．各パラメータ γ，ϕ，θ，α，β，ϕ は順にクラスの出現確率の事前分布のパラメータ，クラスごとのトピックの単語分布，クラスのトピック分布，トピック分布の事前分布のパラメータ，単語分布の事前分布のパラメータである．

トピックごとの単語分布 ϕ とクラスのトピック分布 θ を差分型ギブスサンプリングにより学習し，クラスの出現確率と学習率を γ により推定するが，これらの事前分布であるディリクレ分布のパラメータ α，β，γ の更新も行う．各パラメータの更新はminka[25]による不動転反復法を用いる．

ディリクレ分布はガンマ関数によって決定され，ディガンマ関数はガンマ関数の対数微分によって決定される．

$$\Psi(x) = \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

ディガンマ関数は以下の漸化式を満たす．ここで δ はオイラーの定数である．

$$\Psi(x+1) = \Psi(x) + \frac{1}{x} = -\delta + \sum_{k=1}^{\infty}(\frac{1}{k} - \frac{1}{x+k})$$

以下の式よりディリクレ分布のパラメータである α,β と γ の更新を行う．

$$\alpha_j^{new} = \alpha_j \frac{\sum_d \Psi(C_{j,d}^{doc} + \alpha_j) - \Psi(\alpha_j)}{\sum_d \Psi(\sum_j C_{j,d}^{doc} + \sum_j \alpha_j) - \Psi(\sum_j \alpha_j)} \tag{6.10}$$

$$\beta_m^{new} = \beta_m \frac{\sum_j \Psi(C_{m,j}^{word} + \beta_m) - \Psi(\beta_m)}{\sum_j \Psi(\sum_m C_{m,j}^{word} + \sum_m \beta_m) - \Psi(\sum_m \beta_m)} \tag{6.11}$$

$$\gamma_c^{new} = \gamma_c \frac{\sum_d \Psi(C_{d,c}^{class} + \gamma_c) - \Psi(\gamma_c)}{\sum_d \Psi(\sum_c C_{d,c}^{class} + \sum_c \gamma_c) - \Psi(\sum_c \gamma_c)} \tag{6.12}$$

ここで $C_{doc}^{j,d}$ はトピック j が文書 d に割り当てられた回数，$C_{word}^{m,j}$ は単語 m がトピック j に割り当てられた回数，$C_{class}^{d,c}$ は文書 d がクラス c に割り当てられた回数である．α の更新は文書単位で行い，文書ストリーム中では学習文書に対してのみ行う．β の更新は学習文書の学習後に一度行い，以降は 1 日単位の学習文書に対してのみ行う．γ はストリーム中の学習文書とテスト文書の両方より 1 日単位で行う．

表 6.1: ラベル識別子

Identifier	Corpus Labels
C24	CAPACITY/FACILITIES
C31	MARKETS/MARKETING
C33	CONTRACTS/ORDERS
CCAT	CORPORATE/INDUSTRIAL
GCAT	GOVERNMENT/SOCIAL
GCRIM	CRIME, LAW ENFORCEMENT
GDIP	INTERNATIONAL RELATIONS
GPOL	DOMESTIC POLITICS
GSPO	SPORTS
GVIO	WAR, CIVIL WAR
M11	EQUITY MARKETS
M12	BOND MARKETS
MCAT	MARKETS

表 6.2: Reuter Corpus

class	M11 MCAT	GCAT GSPO	M12 MCAT	GCAT GPOL	GCAT GVIO	
training	802	722	390	269	290	
test	5152	5217	2048	1895	1815	
class	C31 CCAT	GCAT GDIP	C24 CCAT	GCAT GCRIM	C33 CCAT	(Total)
training	230	237	225	269	196	3630
test	1872	1454	1333	1126	1042	22954

6.5 実験

6.5.1 実験準備

実験に用いる Reuters Corpus(RCV1[30]) は 1996 年 8 月 20 日から 1997 年 8 月 19 日までの 810,000 記事の 1 年分のニュース記事からなり，各記事には 128 種類からなるラベルが複数付いている．本実験では，2 つのラベルが付与されている記事を抽出し，その中で頻度が上位である 10 個の組み合わせをクラスとして用いる．用いるラベルとその文書数を表 8.2 に示す．学習データは先頭から 10 日間の 3630 記事であり，テストデータには続いて 60 日間の 22954 記事を使用する．表 8.3 には，コーパスの各識別子とラベル名を示す.

トピックモデルの各パラメータの値はトピック数 10，ギブスサンプリングの繰り返し回数 100 回，差分型ギブスサンプリングの繰り返し回数 20 回，不動点反復法の繰り返しを 5 回とし，事前分布のパラメータの初期値は α =0.5， β =0.01， γ =0.5，学習率の初期値 λ =0.2 とする．また，テスト文書の 1%(230 記事) と 5%(1147 記事) を学習文書として使用する.

実験結果の比較として LDA によって学習されたトピックの確率分布と単語の確率分布を用いた分類手法を使用する．ここではストリーム中で確率分布の更新を行わず，クラスの出現確率も考慮しない．また，文書ストリーム中の学習文書の割合を 0%(テスト文書のみでパラメータ学習)，1% labeled，5% labeled，5% labeled only(学習文書の

表 6.3: KL 情報量

Duration	10-20	10-30	10-40	10-50	10-60	10-70
KL	86691	89224	90480	91539	92022	92419
(increment)	-	+2.92%	+4.37%	+5.59%	+6.15%	+6.61%

みを学習に使用) と変化させ，学習手法の評価を行う．

6.5.2 評価尺度

実験の評価には F 値を用いる．F 値は再現率と適合率の調和平均であり，実際に正解であるもののうち，正解であると予測されたものの割合である再現率 R_i と，正解と予測したデータのうち，実際に正解であるものの割合である適合率 P_i を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \tag{6.13}$$

$$P_i = \frac{a_i}{a_i + b_i} \tag{6.14}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負となる数である．この 2 つの式の調和平均である F 値を次のように定義する．全体の F 値はすべての f_i の平均である．

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad f = Average_i f_i$$

次に，実験データ中の単語分布の変化を測るために分布の差の尺度である KL 情報量を用いる．KL 情報量は 2 つの分布が類似するほど値が小さくなり，KL 情報量の値が大きくなれば分布の異なりが大きいことを表す．学習データの単語分布を V_{train}，テストデータの単語分布を V_{test} としたとき，KL 情報量は以下の式で定義される．

$$KL(V_{train}//V_{test}) = \sum_i V_{train,i} log \frac{V_{train,i}}{V_{test,i}} \tag{6.15}$$

6.5.3 実験結果

実験結果を表 7.3，表 7.5 に示す．表 7.3 より全体の F 値は学習文書 5%使用で 0.883，1%使用で 0.871，テスト文書からのみの学習で 0.870，ベースラインで 0.854 となっている．提案手法ではベースラインと比較して F 値が+1.6 から+2.9 改善されている．こ

のことからオンライントピックモデルを用いた学習手法は有効であると考えられる．最も精度が高くなったのは5%を学習データに用いたときであり，すべてのクラスでベースラインよりF値が高くなっている．提案手法の中で最も悪い0%では1つのクラスで-0.3と下回っているが，その他のクラスではすべてベースラインを上回っている．表7.5より，10日目のf値は提案手法で0.885，比較手法で0.869であり，最終日の累積のF値との差は提案手法の5%で-0.002(0.2%)，ベースラインで-0.015(1.7%)となっており，ベースラインでは提案手法と比較して約8倍悪化している．また，5% labeledと1% labeledのF値から学習データの割合が多い場合により精度が高くなっている．5% labeledと5% labeled onlyの比較からテストデータを学習に使用することでより精度が上昇している．

表6.3より，10日目のKL情報量は86691であり，60日目のKL情報量92419となっており，時間が経過するにつれて学習データとテストデータの単語の確率分布の差が大きくなっている．

6.5.4 考察

分類結果から提案手法のF値は0.870以上であり，オンライントピックモデルによる学習が効果的に働いていると考えられる．学習文書5%と1%を比較すると追加の学習文書を増やすことでF値が1.2%上昇している．テスト文書からのみの学習では学習文書を使用した学習より精度が悪化しているが，ベースラインより精度が1.6%上昇しており，テスト文書からの学習の効果が表れている．学習文書とテスト文書から学習を行う5% labeledと学習文書からのみ学習を行う5% labeled onlyを比較すると，5% labeledのほうがF値が0.3%高くなっており，学習文書とテスト文書の両方を学習に使用することでF値が上昇している．

また，動的学習を行った提案手法では学習文書とテスト文書の確率分布の差が大きくなるのに影響されず，KL情報量が6.61%上昇しても5% labeledでF値が0.2%減少とほとんど変化していない．学習を行わないベースラインではF値が1.6%減少しており，KL情報量が増加するにつれて精度が悪化している．

これらのことから，提案手法では特徴の変化を学習することによりテスト文書の話題の変化に影響されず精度を維持して分類が行え，学習文書とテスト文書の両方を学習に使用することにより高精度で分類が行えることから文書ストリームの分類に有効であると考えられる．

6.6 結び

本研究では，トピックモデルによる分類，文書ストリームに対するオンライントピックモデルの適用，部分学習文書を用いたオンライン分類の3つを行う提案手法を示した．提案手法による分類ではF値が0.883となっており，ベースラインと比較して精

度が上昇することを示した．また，10 日目の F 値が 0.885，60 日目の F 値が 0.883 とほとんど変化しておらず，テスト文書の特徴の変化が大きくなって正解率を維持できることを示した．

表 6.4: 分類の F 値

	M11 MCAT	GCAT GSPO	M12 MCAT	GCAT GPOL	GCAT GVIO	
(5% labelled)						
Recall	0.981	0.964	0.974	0.831	0.923	
Precision	0.982	0.998	0.948	0.878	0.833	
F-value	0.981	0.981	0.961	0.854	0.876	
(vs Baseline)	+1.6	+2.3	+0.5	+3.3	+2.7	
(1% labelled)						
Recall	0.978	0.957	0.973	0.817	0.907	
Precision	0.979	0.998	0.941	0.854	0.802	
F-value	0.979	0.977	0.956	0.835	0.851	
(vs Baseline)	+1.4	+1.9	+0.0	+1.4	+0.2	
(unlabelled)						
Recall	0.977	0.958	0.974	0.823	0.912	
Precision	0.979	0.998	0.938	0.845	0.801	
F-value	0.978	0.977	0.956	0.834	0.853	
(vs Baseline)	+1.3	+1.9	+0.0	+1.3	+0.4	
(5% labelled only)						
Recall	0.987	0.967	0.963	0.850	0.936	
Precision	0.969	0.998	0.957	0.853	0.806	
F-value	0.978	0.982	0.960	0.851	0.866	
(vs Baseline)	+1.3	+2.4	+0.4	+3.0	+1.7	
(Baseline)						
Recall	0.937	0.921	0.973	0.794	0.902	
Precision	0.994	0.999	0.940	0.850	0.802	
F-value	0.965	0.958	0.956	0.821	0.849	
	C31 CCAT	GCAT GDIP	C24 CCAT	GCAT GCRIM	C33 CCAT	(Total)
(5% labelled)						
Recall	0.856	0.836	0.746	0.887	0.893	0.889
Precision	0.906	0.806	0.881	0.873	0.711	0.882
F-value	0.881	0.821	0.808	0.880	0.791	0.883
(vs Baseline)	+2.0	+2.6	+4.1	+2.0	+8.9	+2.9
(1% labelled)						
Recall	0.834	0.807	0.734	0.874	0.892	0.877
Precision	0.895	0.797	0.873	0.859	0.695	0.869
F-value	0.864	0.802	0.798	0.867	0.781	0.871
(vs Baseline)	+0.3	+0.7	+3.1	+0.7	+7.9	+1.7
(unlabelled)						
Recall	0.826	0.794	0.737	0.869	0.888	0.876
Precision	0.894	0.809	0.868	0.856	0.694	0.868
F-value	0.858	0.801	0.798	0.863	0.779	0.870
(vs Baseline)	-0.3	+0.6	+3.1	+0.3	+7.7	+1.6
(5% labelled only)						
Recall	0.843	0.790	0.734	0.878	0.880	0.883
Precision	0.916	0.836	0.890	0.881	0.718	0.882
F-value	0.878	0.813	0.805	0.880	0.791	0.880
(vs Baseline)	+1.7	+1.8	+3.8	+2.0	+8.9	+2.6
(Baseline)						
Recall	0.851	0.830	0.686	0.864	0.943	0.870
Precision	0.870	0.763	0.870	0.856	0.559	0.850
F-value	0.861	0.795	0.767	0.860	0.702	0.854

表 6.5: F 値の推移

days	10	20	30	40	50	60	difference
5% labelled	0.885	0.889	0.888	0.883	0.882	0.883	-0.002 (-0.2%)
1% labelled	0.881	0.884	0.883	0.876	0.872	0.871	-0.010 (-1.1%)
unlabelled	0.881	0.882	0.882	0.874	0.871	0.870	-0.011 (-1.2%)
5% labelled only	0.887	0.886	0.886	0.881	0.879	0.880	-0.007 (-0.8%)
Baseline	0.869	0.868	0.866	0.858	0.855	0.854	-0.015 (-1.6%)

第7章　オンライン学習によるストリーム中の特徴抽出

7.1　前書き

近年，インターネットの発達や電子文書の増大から大量の文書集合を分類する必要性が高まっている．また，ブログ，ツイッター，LINE やフェイスブックに代表されるビッグデータやソーシャルネットワークサービスには多種多様な情報が含まれており，ストリームデータの解析を行うことで現在の流行や評判情報など様々な知識が抽出可能である．特に，ニュース記事，マイクロブログに代表されるストリームデータはリアルタイムな情報を扱うことから，極めて重要な情報源として注目を集めている．このような文書集合ではデータ量が膨大となるため，人手で分類を行うことは困難であり，自動的に分類を行う手法が必要不可欠である．

文書をクラスごとに分類する文書分類では，政治，スポーツ，IT といったクラスは文書の集合として表現され，文書の様々な特徴によって各クラスに分類する．文書の話題は話題独自の特徴によって，固有な単語やそれらの分布を形成する．これにより文書には話題の影響によって特定の単語が多く出現する．また，文書は一般的に異なる特徴を持つ複数の話題によって構成される．

次々に新たなニュースが発生するニュースストリームのような動的な文書集合では，話題の変化に応じて分類先となるクラスの特徴が変化する．例えば，政治に関する話題では，選挙，予算編成，法案審議のような時期に依存したイベントや，外交問題や汚職事件など不定期なイベントに依存し，発生する単語分布は大きく異なる．このためクラスの単語分布はクラス内で共通して現れる特徴と話題による特徴の混合で表される．ニュースストリームの分類ではクラスの特徴を表す定常的な特徴と，該当のクラスで論じられる話題の変化による特徴の変動を加味して分類を行う必要がある．クラスを表す定常的な特徴は，クラス内で話題によらず一定して現れる特徴であり，該当のクラスに属する全ての文書から特徴を学習することが可能である．話題による特徴の変動は，時期やイベントに依存することから，定常的な特徴とは異なり現在の状態に合わせて特徴を学習する必要がある．この 2 つの特徴は異なる分布を持つことから同一の確率分布で扱うことは困難であり，別々の確率分布としてモデル化することでより高精度に分類を行える可能性がある．

本研究では，ニュースストリーム中の文書を動的にカテゴリ分類するため，2 つの問題を論じる．第 1 の問題は，クラスを表す定常的な特徴と話題による変動を如何に学

習するか，第 2 の問題は，特徴の変化が起きるニュースストリームで各話題に対する適切な特徴を如何に学習するかである．トピックモデルは確率論の観点から文書集合のモデル化を行う手法であり，著者推定 [?] や文脈を考慮した情報検索 [36] など様々な目的に用いられている．このモデルでは文書をトピックの混合として，各トピックを単語についての確率分布として表現する．これらの分布はディリクレ事前分布によって制御する．クラスが複数の話題を持つ場合，単一の単語分布で表現することは困難であるが，トピックモデルではクラスをトピックの混合として表現できるため適している．特にトピックモデルのオンライン学習により，話題の変化に応じて新たな特徴を捉えることができる．

本稿では，トピックモデルに基づく提案手法により，ニュースストリーム中で動的に文書分類を行う方式を提案する．提案手法では，クラスの定常的な特徴とストリーム中に発生する局所的な変動を 2 つの種類のトピック分布として表すことで分類を行う．ここでは，週，月や季節といった一定の区間ごとに変化する確率分布を考えるのではなく，直近の文書に対してウィンドウを設定し，このウィンドウを遷移させていくことで変動する特徴を取り込む．

第 7.2 章では関連研究について述べ，第 7.3 章ではニュースストリームについて述べる．第 7.4 章ではトピックモデルについて述べ，第 7.5 章では提案手法について述べる．第 7.6 章では実験により有効性を示す．第 7.7 章で結びとする．

7.2 関連研究

近年トピックモデルを用いた多くの研究が行われ様々な応用例が示されている．Ramage らは学習文書を使用した教師有りのトピックモデルとして，トピックとクラスを 1 対 1 で対応させることで LDA に著者の情報を取り入れた Labeled LDA を提案している [28]．Rosen-Zvi らによる Author-Topic(AT) モデルでは，著者ごとにトピック分布を持たせることで文書単位ではなく著者単位のトピック分布を推定している [31]．ここでは，文書内の著者とトピックの偏りを考慮していない．また，トピックモデルでは時系列データを扱うための拡張も多数提案されている．Wang らによる Topics over Time(TOT) モデルでは LDA にタイムスタンプを取り入れることで，トピックがタイムスタンプの分布を持つ．これにより，時系列データ内でのトピックの出現の変化を考慮することができる [44]．Naveed らによる Author-Topic-Time(ATT) モデルは AT モデルと TOT モデルの発展形であり，AT モデルにタイムスタンプを取り入れている [27]．このモデルではトピックと著者がタイムスタンプによる時間分布を持つことで，時間によるトピックと著者の出現の変化を考慮している．これにより文書が書かれた時期によるトピック分布と文書内での著者の分布を捉えることができる．ATT モデルは観測された時系列データに対するモデル化であり，新たに到着した文書を分類するストリームの分類には適していない．Iwata らによる Online Multiscale Dynamic Topic モデルでは，異なるスケールでの単語の依存性を考慮することで，複数のスケールでト

ピックの時間発展を捉えることができる [39]．また，Li らによる Pachinco Allocation モデルのようにトピックに階層構造を持たせたトピックモデルも提案されている [20]．ここでは，有向非循環グラフを用いてトピック構造を推定することで，トピック同士の関係を捉えることができる．トピックに階層構造を取り入れることでより高精度に文書集合のモデル化を行える可能性があるが，ストリーム中では階層構造も変化していくことが考えられるので本稿ではトピックの階層構造を扱わない．

トピックモデルのオンライン学習 [14, 32] では，差分型ギブスサンプリングや EM アルゴリズムによるパラメータ学習により，一括学習の適用が困難な膨大な文書集合を効率よく扱う．また，ストリームデータなどの動的な文書集合に適用すれば，集合の変化に応じてパラメータを更新することで，新たな特徴を学習することができる．Canini らはギブスサンプリングを基にした差分型ギブスサンプリングを提案している [7]．差分型ギブスサンプリングは，現在までに得られたパラメータを基に各単語にトピックを割り当て，その後活性化ステップにおいて過去のトピック割り当てをランダムに選択し，再サンプリングを行う．Yao らは教師データから学習したパラメータを基に新たに到着した文書内で繰り返し学習を行うギブスサンプリング手法により，文書分類が行えることを示している [50]．しかし，ここではトピックモデルのパラメータを更新するのみであり，ストリームデータの特性を考慮していない．AlSumait らは時刻 t でのトピックごとの単語分布に時刻 $t-1$ での出現確率を事前分布として用いることで，それ以前のデータを必要としない LDA のオンライン学習手法である OLDA を提案している [1]．

本稿のように異なる性質を持つトピックを扱うトピックモデルとして，Kawamae らは，タイムスタンプが与えられた文書集合に対して，定常トピックと時制トピックを検出する Theme Chronicle Model を提案している [17]．TCM は，スイッチ変数により各単語を出力するトピックを選択することで，話題，トレンド，文書依存，バックグラウンドという 4 種類のトピックを学習する．話題は文書ごとにトレンドの分布を持ち，タイムスタンプによってトレンドの出現が変化する．しかしながら，ここではクラスのトピック分布やその変化を直接的には考慮していない．Ida らは，オンラインレビューのような数値レートを持つ文書集合に対するモデル化として，領域依存と領域独立のトピックを用いた Domain-Dependent/Independent Topic Switching Model を提案している [38]．DDITSM は，文書の領域と数値レートを教師情報として持つため，数値レートにより領域を分割でき，さらに領域依存トピックとして領域の特徴を抽出することができる．ここでは，Labeled LDA と同様に領域とトピックが 1 対 1 で対応しており，領域が複数の領域依存トピックを持つことを考慮していない．

7.3 ニュースストリーム

ニュースストリーム，マイクロブログに代表される SNS データは，極めて重要な情報源として注目されており，娯楽，広告や企業分析等の目的で多数の人々が利用して

いる．ニュースストリームでは，各記事は記事が属するクラスが持つ複数の話題を含み，話題に関連して出現する単語分布が変化する．この話題はストリーム中で大きく変化する可能性がある．また，一度新たな話題が発生すると関連した記事が短期間に多数発生し，単語分布が大きく変わるというバースト現象が発生する [18]．各話題に依存した単語分布は，新たなイベントの影響を強く受ける．例えば，オリンピックのような大きなイベントが起これば，そのイベントに注目が集まるため期間中はオリンピックに関する記事が急増する．期間が終わると異なる話題に遷移していきオリンピックに関する記事は減少する．この話題の変化によってクラスの特徴が変化していくため，動的な学習により分類基準を変更する必要がある．

7.3.1 問題設定

本研究で扱うニュースストリームの分類を以下のように設定する．ニュースストリームはラベル付き文書 $K = k_1, k_2, ..., k_i$ とラベル無しの文書 $U = u_1, u_2, ..., u_j$ の混合で構成される．分類では，初期の学習文書集合から各クラスの特徴を学習し，ストリーム中で新たに到着するラベル無しの文書 u のクラスを推定する．また，到着した文書がラベル付き文書 k である場合は追加の学習を行う．

7.4 トピックモデル

本研究ではトピックモデルに基づくニュースストリームに対する新たな分類手法を提案する．トピックモデルは文書集合の分析に非常に有効なモデルとして注目されており，文書分類や文書要約など幅広く用いられる [37]．トピックモデルでは，文書をトピックの混合分布として複数の確率分布で表す．これにより文書の持つ複数の話題が表現可能となる．LDA(Latent Dirichlet Allocation) はトピックモデルの代表例として多数利用されている [5]．確率的潜在意味索引付け (Probabilistic Latent Semantic Indexing, pLSI) は, LDA と違って, トピックの混合比を学習データの文書集合に依存して固定化している [13]. 一方, LDA ではこの混合比は事前分布から生成する点で異なる. LDA は学習データだけに依存しないが, 代わりに特定の確率モデルを仮定するため, 柔軟な観点で文書を扱える可能性がある.

トピックモデルをニュース記事集合に適用することで，クラスをトピックの分布で表し，クラス内の複数の話題を特徴付けることが可能となる．話題は記事の内容を意味し，トピックは特定の意味を持たない単語のクラスタであるが，話題はトピックに対応すると仮定する．これにより話題の変化をトピックの変化として表現する．

2 章図 2.1(a) は LDA のグラフィカルモデルである．図中の変数は, 図 2.1(a) 左に, ディリクレ事前分布 $Dir(\beta)$，図 2.1(a) 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1(a) 左上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1(a) 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N_d は各文書の単語数を表す．LDA

の単語生成過程を以下で示す．まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する．次に, 文書 d 内の i 番目の単語 w_i において，抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する．

LDA の拡張である AT モデルでは，2 章図 2.1(b) で表される教師情報 $\mathbf{a_d}$ により文書の著者など該当の文書が属するクラスの情報を持つ．文書 d において $\mathbf{a_d}$ から著者 x は一様に得られ，トピック z はクラスごとのトピック分布に基づきトピックから単語が生成される．LDA は文書ごとにトピック分布を持つが，AT モデルでは著者ごとにトピック分布を持つ点で異なる．

これらのモデルは一般的に静的な文書集合に対して適用されるが，LDA をストリームに対応させるためにオンライン学習手法が示されている [2]．Banerjee らによるオンライン学習では，まず現在までの文書集合に対して，パラメータを推定する．得られたパラメータを基に新たな文書のパラメータを MAP 推定することでパラメータを更新する．

7.5 提案手法

7.5.1 提案モデル

本稿は，トピックモデルに基づくニュースストリームに対する分類手法を提案する．提案モデルは，AT モデルと同様に文書のクラスラベルを教師情報として使用することで，クラスごとのトピック分布を学習する．さらに，定常トピックと変動トピックという 2 種類のトピックを用いることで，ストリーム中でのトピックの特徴の変化を捉える．定常トピック分布は，これまでに出現した全ての文書から得られるクラスの特徴を表し，変動トピック分布はウィンドウにより直近の文書の特徴を表す．ニュースストリームでは特徴の変動が不定期に起きる可能性があるため，タイムスタンプにより日や週といった区間ごとに変化する確率分布を用いるのではなく，直近の文書に対してウィンドウを遷移させていくことで特徴の変化を考慮する．

図 7.1 は提案モデルのグラフィカルモデルである．図中の変数は, 図 7.1 左に, ディリクレ事前分布 $Dir(\alpha)$，各クラスの定常トピック分布 θ^{st}，L はクラス数，図 7.1 左下に，ディリクレ事前分布 $Dir(\beta)$，単語空間の多項分布 $Multinomial(\phi^{st}_{z_i^{st}})$, T はトピック数, 図 7.1 中央にベータ事前分布 $B(\gamma)$, 文書のクラス c, 二項分布 $Binomial(\lambda)$, x は指示変数であり単語 w_i に定常トピック st と変動トピック tr のどちらが割り当てられるかを表す．図 7.1 右に各クラスの変動トピック分布 θ^{tr}，多項分布 $Multinomial(\phi^{tr}_{z_i^{tr}})$, それぞれのディリクレ事前分布 $Dir(\alpha)$，$Dir(\beta)$，D は文書数, N_d は各文書の単語数を表す．提案モデルの単語生成過程を以下で示す．まず, ベータ事前分布 $B(\gamma)$ から文書のクラス c に関する λ_d を抽出し，文書 d 内の i 番目の単語 w_i が，定常トピックか変動トピックのどちらから生成されるかを選択する．次に選択された x のディリクレ

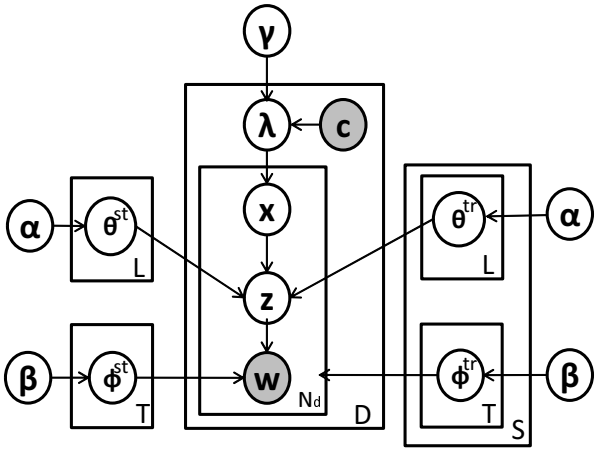

図 7.1: 提案モデル

事前分布 $Dir(\alpha)$ からクラスのトピック分布である θ_l^x を抽出する．抽出したクラス l の多項分布 $Multinomial(\theta_l^x)$ からトピック z_i^x を抽出し，そのトピック z_i^x の多項分布 $Multinomial(\phi_{z_i^x})$ から単語 w_i を抽出する．変動トピック分布は局所的な特徴の変動を表す分布であり，定常トピック分布がすべての文書集合から学習されるのに対し，変動トピック分布ではウィンドウ s 内で学習を行う．ウィンドウ s の大きさを F とした場合，ウィンドウ s は直近のラベル付き文書である $k_{i-F}, k_{i-F+1}, ..., k_i$ で構成される．このウィンドウは新たなラベル付き文書が到着するたびに遷移する．

ギブスサンプリングによりトピックごとの単語分布 ϕ とクラスのトピック分布 θ の値は以下の式で求める．

$$
\begin{aligned}
&P(z_i = j, x \in \{(st), (tr, s)\} | w_i = m, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \\
&\propto \frac{N^x_{m,j} + \beta}{\sum_m N^x_{m,j} + V\beta} \frac{N^x_{c,j} + \alpha}{\sum_j N^x_{c,j} + T\alpha} \frac{N^x_d + \gamma}{N_d + 2\gamma}
\end{aligned} \tag{7.1}
$$

$$
\phi^x_{m,j} = \frac{N^x_{m,j} + \beta}{\sum_m N^x_{m,j} + V\beta} \tag{7.2}
$$

$$
\theta^x_{c,j} = \frac{N^x_{c,j} + \alpha}{\sum_j N^x_{c,j} + T\alpha} \tag{7.3}
$$

$$
x \in \{(st), (tr, s)\}
$$

$N^{st}_{m,j}$, $N^{st}_{c,j}$ は定常トピック分布から単語 m がトピック j に割り当てられた回数，クラス c にトピック j を割り当てた回数，$N^{tr,s}_{m,j}$, $N^{tr,s}_{c,j}$ は変動トピック分布からウィンド

ウ s 内で，単語 m がトピック j に割り当てられた回数，クラス c にトピック j を割り当てた回数，N_d^{st} は文書 d で定常トピックに割り当てられた回数, $N_d^{tr,s}$ は文書 d で変動トピックに割り当てられた回数, V は全単語数，T は全トピック数である．

7.5.2 分類とパラメータの更新

ニュースストリーム中の文書を分類するため，新たに到着した未知文書に対して，各クラスから生成される尤度を求める．分類には提案モデルによって学習した各クラスのパラメータを用い，最も尤度の高いクラスに分類する．尤度は以下の式より求める．

$$Ans(c) = arg \max_c P(d|\phi^{st}, \phi^{tr,s}, \theta_c^{st}, \theta_c^{tr,s}) \tag{7.4}$$

各パラメータ ϕ^{st}, $\phi^{tr,s}$, θ_c^{st}, $\theta_c^{tr,s}$ は順に各クラスの定常トピックの単語分布，各クラスの変動トピックの単語分布, 各クラスの定常トピック分布，各クラスの変動トピック分布である．

ストリーム中では，Banerjee らによるオンライン学習 [2] と同様に新たに到着するラベル有り文書 $\{k_{m+1}, ..., k_{|K|}\}$ を用いて，逐次的にオンライン学習を行う．初期学習では m 個の学習文書 $\{k_1, ..., k_m\}$ 内でギブスサンプリングにより繰り返し学習を行うのに対し，追加の学習では新たに到着した文書のパラメータのみ学習する．新たに到着する文書 d_i が，ラベルを有する $d_i \in K$ のとき追加の学習文書としてオンライン学習を適用する．

7.6 実験

実験では，Reuters Corpus のニュース記事を分類することで，分類精度とストリーム中での精度の変化を比較する．

7.6.1 実験準備

実験に用いる Reuters Corpus(RCV1[30]) は 1996 年 9 月 1 日から 1997 年 8 月 19 日までの 810,000 記事の 1 年分のニュース記事からなり，各記事には 128 種類からなるラベルが複数付いている．本実験では，1996 年 9 月から 1997 年 8 月の各月の先頭 10 日間である計 120 日間で，2 つのラベルが付与されている記事を抽出し，その中で頻度が上位である 10 個の組み合わせをクラスとして用いる．実験データ中の不要語は取り除き，算用数字は*に置き換え長さにのみ着目する．文字は全て小文字に置き換える．用いるラベルとその文書数を表 8.2 に示す．学習データは先頭から 3000 記事であり，テストデータには残りの 48080 記事を使用する．テストデータの中の 10%を等間隔で到

表 7.1: ラベル識別子

Identifier	Corpus Labels
C24	CAPACITY/FACILITIES
C31	MARKETS/MARKETING
C33	CONTRACTS/ORDERS
CCAT	CORPORATE/INDUSTRIAL
GCAT	GOVERNMENT/SOCIAL
GCRIM	CRIME, LAW ENFORCEMENT
GDIP	INTERNATIONAL RELATIONS
GPOL	DOMESTIC POLITICS
GSPO	SPORTS
GVIO	WAR, CIVIL WAR
M11	EQUITY MARKETS
M12	BOND MARKETS
MCAT	MARKETS

表 7.2: ロイターコーパス

class	M11 MCAT	GCAT GSPO	M12 MCAT	C31 CCAT	GCAT GPOL	
training	634	702	234	319	246	
test	11434	11120	5037	4861	3638	
total	12068	11822	5271	5180	3884	
class	GCAT GVIO	GCAT GDIP	C24 CCAT	C33 CCAT	GCAT GCRIM	(Total)
training	264	185	136	150	131	3000
test	3136	2663	2284	2008	1899	48080
total	3440	2848	2420	2158	2030	51080

着するラベル付き文書とし，分類後に文書のラベルを使用してパラメータの更新に用いる．表 7.1 には，コーパスの各識別子とラベル名を示す.

提案手法の各パラメータの値はトピック数 50，変動トピックのウィンドウサイズを 3000 文書, 初期学習のギブスサンプリングの繰り返し回数 200 回，事前分布のパラメータは α =0.5， β =1/語彙数， γ =0.1 とする．

実験の比較手法として，定常トピック，変動トピックと教師情報の組み合わせによる有効性を検証するため，提案モデルに教師情報を用いない簡略形を用いる．簡略形は教師情報を用いないため，クラスごとのトピック分布ではなく，LDA 方式により文書ごとに定常トピック分布と変動トピック分布を推定する．学習後に文書のクラスラベルを用いてクラスごとにトピックの割り当てを合計することで各クラスのトピック分布を得る．簡略形においても提案手法と同様にオンライン学習によりパラメータを更新する．また，他のトピックモデルとの比較として AT モデルと OLDA を使用する．AT モデルによる分類ではストリーム中で確率分布の更新を行わない場合と提案手法と同様のオンライン学習によりパラメータを更新する場合の 2 通りを用いる．OLDA では，テスト中に 3000 文書区切りで学習文書と該当の区間のテスト文書を用いて学習を行いパラメータを更新する．簡略形のトピック数 90，AT モデルのトピック数は 90, OLDA のトピック数は 100 とする．

7.6.2 評価尺度

実験の評価にはF値を用いる．F値は再現率と適合率の調和平均であり，実際に正解であるもののうち，正解であると予測されたものの割合である再現率R_iと，正解と予測したデータのうち，実際に正解であるものの割合である適合率P_iを次のように定義する．

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{7.5}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{7.6}$$

TP_iは推定結果が正である数，FN_iは正であるが負と推定された数，FP_iは正であると推定した中で正解が負となる数である．この2つの式の調和平均である各クラスのF値を次のように定義する．

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

全体の評価としてF値のマイクロ平均であるmicro-Fとマクロ平均であるmacro-Fの2つを用いる．micro-Fは全てのクラスをまとめた再現率と適合率から算出され，macro-Fは各クラスについての再現率と適合率の平均から算出する．

また，トピックが持つ単語分布の変化を比較するのにJSダイバージェンスを用いる．JSダイバージェンスは対称な確率分布の差の尺度であり確率分布Pと確率分布Qが与えられたとき，JSダイバージェンスは以下の式で求まる．

$$JS(P//Q) = \frac{1}{2}(\sum_x P(x) log \frac{P(x)}{R_(x)} + \sum_x Q(x) log \frac{Q(x)}{R(x)}) \tag{7.7}$$

ここでRは確率分布PとQの平均であり，$R = \frac{P+Q}{2}$となる．JSダイバージェンスの値はP=Qのとき0となり，確率分布の差が大きいほど値が大きくなる．

7.6.3 実験結果

実験結果を表7.3，表8.9, 表7.5に示す．表7.3より全体の各F値は提案手法で0.875, 0.919，簡略形で0.794，0.850，OLDAで0.811，0.877，AT更新で0.825，0.880，ATで0.806，0.865となっている．提案手法と簡略形の比較では，学習時に教師情報を組み込むことでmacro-F値で8.1%，micro-F値で6.9%改善している．提案手法とOLDAの比較ではmacro-F値で6.4%，micro-F値で4.2%改善している．提案手法とATモデルの比較では，macro-F値で5%，micro-F値で3.9%改善している．ATモデルは更新を行うことでF値がそれぞれで1.9%と1.5%上昇している．表8.9のトピック数別のF値

では，提案手法はトピック数 20 からトピック数 100 の全てで最も F 値が高くなっている．OLDA ではトピック数が増えるにつれて精度が上昇しているが，トピック数 90 と 100 では macro-F 値で 0.0%，micro-F 値で 0.1%上昇とほとんど変化しなくなっている．

表 7.5 より累計の F 値の推移では，提案手法の macro-F 値は 10000 文書目と合計で共に 0.875 と変化していないが，AT 更新と AT ではそれぞれ-0.3%，-1.7%と減少している．micro-F 値はパラメータ更新を行う提案手法と AT 更新ではそれぞれ+0.9%，+0.6%と上昇しているが，パラメータの更新を行わない AT では-0.4%と F 値が減少している．OLDA では 10000 文書目が macro-F 値で 0.787，micro-F 値で 0.847 と特に低くなっていることから F 値の上昇幅は+2.4%と+3.0%と最も大きくなっている．簡略形では macro-F 値と micro-F 値がそれぞれ+0.4%，+1.2%上昇している．

表 7.6，表 7.7 は，分類開始時，15000 文書目，30000 文書目，45000 文書目での定常トピックと変動トピックで確率が上位となる語である．表 8.12 は前後の時点でのトピックの単語分布を比較した JS ダイバージェンスである．

7.6.4 考察

分類結果から提案手法の F 値は 0.875，0.919 と最も高くなっている．簡略形に対して macro-F 値で 8.1%，micro-F 値で 6.9%改善しており，文書の教師情報を学習時に使用してモデルの学習を行うことで分類精度が向上している．各クラスの F 値では 10 クラス中 8 クラスで提案手法の F 値が高くなっている．簡略形では初期の学習文書数が最も多い”GCAT GSPO”クラスで特に高精度になっており，学習文書数の少ないクラスでは精度が低くなっている．提案手法ではクラスごとにトピック分布を持っており，文書のクラス情報を使用して学習を行うことから学習文書数の少ない文書でも高精度に分類が行えている．AT モデルでは追加の学習データを用いてパラメータ更新を行うことで，学習を行わない場合と比較して精度が上昇しているが，macro-F 値では 10000 文書目と合計の差が-0.3%と精度が減少している．micro-F 値では+0.6%と上昇しているため，テスト文書数の少ないクラスで誤分類が増えたために精度が悪化したと考えられる．文書数の少ないクラスでは新たな学習文書も少なくなるため変化に対応することが困難になっている．提案手法では，直近の文書で構成する変動トピックを用いているため，macro-F 値においても精度を維持して分類が行えている．また，定常トピックと変動トピックの各トピックで確率が上位となる単語を見てみると，定常トピックでは 4 つの時点のトップ 5 である 20 語の単語の種類数は各トピックで順に 8，9，8 と多くの単語が重複して出現している．変動トピックでは各トピックで順に 9，11，11 と定常トピックと比較してトピック内で確率の高くなる単語が多く変化している．表 8.12 より，提案手法の定常トピックと AT モデルでは学習文書数が増えるにつれてトピック内の変動が単調に減少しているが，変動トピック分布は JS ダイバージェンスが単調に減少せずに増減しているためトピックの中身が変化している．これにより，提案手法と AT 更新の各 F 値の差は 10000 文書目で 3.6%，4.7%であったのに対し，合計では 3.9%と 5.0%と精度の差が大きくなっている．

提案手法では特徴の変化を学習することによりテスト文書の話題の変化に影響されず精度を維持して分類が行え，新たなラベル付き文書を学習に使用することにより高精度で分類が行えることからニュースストリームの分類に有効であると考えられる．

7.7 結び

本研究では，定常トピック分布と変動トピック分布によりクラスの特徴を学習し，新たに到着したラベル付き文書を用いてパラメータを更新する分類手法を示した．提案手法による分類では macro-F 値で 0.875，micro-F 値で 0.919 となっており，簡略形，AT モデル，OLDA と比較して精度が上昇することを示した．また，10000 文書目の macro-F 値が 0.875，最終の F 値が 0.875 と変化しておらず，分類を継続しても正解率を維持できることを示した．

表 7.3: 各クラスの F 値

	M11 MCAT	GCAT GSPO	M12 MCAT	GCAT GPOL	GCAT GVIO	
(提案手法)						
Recall	0.981	0.867	0.824	0.847	0.878	
Precision	0.994	0.927	0.872	0.755	0.798	
F-value	0.988	0.896	0.848	0.798	0.836	
(簡略形)						
Recall	0.956	0.933	0.815	0.805	0.709	
Precision	0.921	0.998	0.918	0.740	0.835	
F-value	0.938	0.964	0.863	0.771	0.767	
(OLDA)						
Recall	0.981	0.870	0.842	0.650	0.827	
Precision	0.971	0.846	0.816	0.636	0.756	
F-value	0.976	0.858	0.829	0.643	0.790	
(AT+更新)						
Recall	0.953	0.817	0.677	0.818	0.882	
Precision	0.997	0.914	0.886	0.642	0.708	
F-value	0.974	0.863	0.768	0.720	0.786	
(AT)						
Recall	0.949	0.802	0.565	0.777	0.902	
Precision	0.997	0.885	0.901	0.656	0.649	
F-value	0.972	0.841	0.694	0.712	0.755	
	C31 CCAT	GCAT GDIP	C24 CCAT	GCAT GCRIM	C33 CCAT	macro-F
(提案手法)						
Recall	0.883	0.766	0.950	0.987	0.776	0.876
Precision	0.826	0.811	0.966	0.965	0.829	0.874
F-value	0.854	0.788	0.958	0.976	0.802	0.875
(簡略形)						
Recall	0.780	0.714	0.550	0.880	0.834	0.798
Precision	0.818	0.645	0.777	0.719	0.525	0.790
F-value	0.799	0.678	0.644	0.791	0.644	0.794
(OLDA)						
Recall	0.770	0.637	0.887	0.958	0.647	0.807
Precision	0.790	0.689	0.914	0.954	0.769	0.814
F-value	0.780	0.662	0.900	0.956	0.703	0.811
(AT+更新)						
Recall	0.845	0.723	0.901	0.972	0.706	0.829
Precision	0.754	0.705	0.930	0.958	0.706	0.820
F-value	0.797	0.714	0.915	0.965	0.706	0.825
(AT)						
Recall	0.828	0.766	0.867	0.968	0.656	0.808
Precision	0.742	0.639	0.936	0.953	0.685	0.804
F-value	0.783	0.697	0.900	0.961	0.670	0.806

表 7.4: トピック数別の F 値

トピック数	20	30	40	50	60	70	80	90	100
(提案手法)									
macro-F	0.769	0.841	0.854	0.875	0.849	0.851	0.815	0.868	0.859
micro-F	0.817	0.898	0.906	0.919	0.903	0.898	0.866	0.915	0.910
(簡略形)									
macro-F	0.677	0.711	0.716	0.762	0.772	0.768	0.782	0.794	0.770
micro-F	0.747	0.764	0.776	0.811	0.828	0.821	0.830	0.850	0.826
(OLDA)									
macro-F	0.743	0.775	0.780	0.794	0.790	0.791	0.794	0.811	0.811
micro-F	0.823	0.848	0.854	0.865	0.860	0.863	0.865	0.876	0.877
(AT 更新)									
macro-F	0.741	0.776	0.795	0.794	0.797	0.809	0.800	0.825	0.820
micro-F	0.813	0.844	0.856	0.857	0.861	0.869	0.862	0.880	0.878
(AT)									
macro-F	0.728	0.766	0.775	0.776	0.775	0.787	0.782	0.806	0.801
micro-F	0.800	0.833	0.839	0.842	0.843	0.851	0.848	0.865	0.863

表 7.5: F 値の推移

文書数	10000	20000	30000	40000	合計	差
(提案手法)						
macro-F	0.875	0.876	0.876	0.877	0.875	0.000 (0.0%)
micro-F	0.910	0.913	0.914	0.918	0.919	+0.009 (+0.9%)
(簡略形)						
macro-F	0.790	0.794	0.790	0.793	0.794	0.004 (0.4%)
micro-F	0.838	0.843	0.841	0.848	0.850	+0.012 (+1.2%)
(OLDA)						
macro-F	0.787	0.799	0.806	0.809	0.811	+0.024 (+2.4%)
micro-F	0.847	0.860	0.867	0.873	0.877	+0.030 (+3.0%)
(AT 更新)						
macro-F	0.828	0.827	0.823	0.825	0.825	-0.003 (-0.3%)
micro-F	0.874	0.875	0.873	0.878	0.880	+0.006 (+0.6%)
(AT)						
macro-F	0.823	0.814	0.807	0.807	0.806	-0.017 (-1.7%)
micro-F	0.869	0.864	0.858	0.863	0.865	-0.004 (-0.4%)

表 7.6: 変動トピック内の単語分布の推移

トピック番号 0							
0		15000		30000		45000	
単語	確率	単語	確率	単語	確率	単語	確率
malaysia	0.023	malaysia	0.023	yugoslavia	0.024	yugoslavia	0.023
taiwanese	0.022	taiwanese	0.015	december	0.013	mohammad	0.022
arrived	0.016	yugoslavia	0.014	leaving	0.013	leaving	0.016
diplomats	0.016	klaus	0.014	arrived	0.012	defeat	0.012
december	0.012	december	0.013	compromise	0.012	arrived	0.012
トピック番号 10							
0		15000		30000		45000	
単語	確率	単語	確率	単語	確率	単語	確率
valery	0.042	burma	0.040	gowda	0.073	gowda	0.065
howard's	0.025	valery	0.037	sese	0.053	sese	0.051
longer	0.025	longer	0.032	seko	0.051	longer	0.048
italians	0.025	prd	0.027	longer	0.042	seguin	0.046
perot	0.020	matthew	0.022	burma	0.032	appointmer	0.046
トピック番号 20							
0		15000		30000		45000	
単語	確率	単語	確率	単語	確率	単語	確率
loader	0.067	saudi	0.095	saudi	0.091	saudi	0.048
iata	0.047	toy	0.038	toy	0.038	hermis	0.045
tyrrell	0.028	sulphur	0.035	galatasaray	0.035	inflows	0.032
mccallion	0.028	nfp	0.025	sulphur	0.032	galatasaray	0.029
sheikh	0.024	tyrrell	0.022	*nb	0.029	blauensteiner	0.026

表 7.7: 定常トピック内の単語分布の推移

トピック番号 0							
0		15000		30000		45000	
単語	確率	単語	確率	単語	確率	単語	確率
grand	0.021	top	0.019	top	0.018	top	0.017
championship	0.019	championship	0.017	championship	0.017	round	0.017
top	0.019	grand	0.016	soccer	0.014	championship	0.016
soccer	0.014	soccer	0.015	group	0.014	germany	0.014
former	0.013	group	0.013	grand	0.014	group	0.014
トピック番号 10							
0		15000		30000		45000	
単語	確率	単語	確率	単語	確率	単語	確率
issue	0.061	abb	0.059	abb	0.052	abb	0.058
mcveigh	0.051	waste	0.039	issue	0.036	horan	0.039
abb	0.040	issue	0.039	horan	0.032	issue	0.039
crude	0.040	horan	0.034	waste	0.032	waste	0.029
stalexport	0.030	cuellar	0.034	icici	0.028	cuellar	0.026
トピック番号 20							
0		15000		30000		45000	
単語	確率	単語	確率	単語	確率	単語	確率
seasonal	0.059	fair	0.051	fair	0.071	fair	0.071
fair	0.059	seasonal	0.047	queensland	0.051	seasonal	0.052
queensland	0.053	queensland	0.047	seasonal	0.041	queensland	0.042
cheug	0.039	kyi's	0.029	dominican	0.025	cheung	0.029
campaign	0.039	campaign	0.025	belarus	0.023	belarus	0.025

表 7.8: JS ダイバージェンス

文書番号	0 ↔ 10000	10000 ↔ 20000	20000 ↔30000	30000↔ 40000
定常トピック	1.93	1.23	0.887	0.647
変動トピック	6.06	4.76	5.60	4.74
AT 更新	0.812	0.553	0.422	0.289

第8章　ストリーム中の複数のラベルを持つ文書からの特徴抽出

8.1　前書き

近年，マイクロブログ，Facebook，ニュース記事に代表される文書ストリームデータは極めて重要な情報源として注目されている．文書を単位とするストリームデータではタイムリな情報を扱うことから，文書集合の解析によって，現実の出来事に即した現在の流行や評判情報など様々な知識が得られる．

文書をラベルごとに分類する文書分類では，政治，スポーツ，IT といったラベルは文書の集合として表現され，文書の様々な特徴によって各ラベルに分類される．話題は，文書で論じられる内容を表す．同じ話題の文書には話題独自の特徴によって特定の単語が多く出現する．また，マルチラベル分類では，1 つあるいは複数のラベルを持つ文書を仮定する．このラベル間には”経済”と”市場”は共起しやすいが”芸術”と”健康”は共起しにくいといった依存性があることが考えられる．このため，マルチラベル分類では，ラベル間の関係を考慮することが重要となる．

ニュース記事やマイクロブログといった文書ストリームでは，話題の変化に応じて文書集合の特徴が変化するため，分類基準をあらかじめ固定することは困難となる．ニュース記事では新しいニュースが逐次発生するため，話題が動的に変化する．例えば，政治に関する話題では，選挙，予算編成，法案審議のように時期による出来事や，外交問題や汚職事件など不定期な出来事に依存し，発生する単語は大きく異なる．これらの出来事の中でも，時間経過によって論じられる内容は逐次変化している．ラベルには大きく 2 つの特徴がある．時間に影響されずラベル内で共通して表れる定常的な影響と，ラベルで論じられる話題の変化による特徴の変動である．ストリームの文書分類では，特徴の変化に対応するために，ストリーム中で新たなラベル有り文書により追加の学習を行う手法が示されている [46, 43]．本研究においてもストリーム中で新たな学習文書であるラベル有り文書が得られることを仮定する．

本稿では，ラベルごとに共通して表れる話題を定常分布，局所的に出現する話題を変動分布で表し，分類に用いる．この 2 つの分布のモデル化にトピックモデルを用いる．トピックモデルは文書集合の分析に有効なモデルとして文書分類 [34] など幅広く用いられている．トピックモデルでは，文書単位で文書が属するラベルを考えるのではなく，文書内の各単語に潜在状態であるトピックを仮定し，単語にトピックを 1 対 1 で割り当てる．文書はトピックの混合で表される．文書内の複数の話題を表すことが

容易になる．トピックは意味が明示されない単語のクラスタであるが，トピックが話題に対応すると仮定する．これにより，話題の変化をトピックの変化として表現する．特にトピックモデルのオンライン学習により，話題の変化に応じて新たな特徴を捉えることができる．

本論文の貢献は，ストリーム中でマルチラベル分類を行うために 2 つの問題を論じることにある．第 1 の問題は，複数のラベルを許す文書ストリームでのラベルの学習方法，第 2 の問題はラベル間の共起関係を含むラベルの推定方法を扱う．本論文の結果は主に以下の 2 点である．(1) ストリーム中の話題の変動を考慮したトピックモデルに基き，マルチラベルを持つ文書のモデル化が可能である．(2) ラベルの共起関係を考慮したラベリング手法を提案する．

本稿は，複数のラベルを持つ文書集合から各ラベルに対応する文書集合の特徴を動的に学習する方式に基く．テスト文書のラベリングは，推定結果となるマルチラベルを直近のラベル有り文書で構成されるウィンドウ内で出現したマルチラベルに限定する．これにより，高精度に文書ストリームのマルチラベル分類が行える．

第 8.2 章では関連研究について述べ，第 8.3 章では文書ストリームのマルチラベル分類について述べる．第 8.4 章ではトピックモデルについて述べ，第 8.5 章では提案手法について述べる．第 8.6 章では実験により有効性を示し，第 8.7 章で結びとする．

8.2 関連研究

文書ストリームでは集合内の特徴が動的に変化するため，ストリームの特性を考慮した解析手法が必要となる．特にニュース記事のようなストリームでは一旦新たな話題が発生すると関連した記事が短期間に多数発生し，単語分布が劇的に変化するというバースト現象が発生する [18]．また，これに関連してストリーム中では時間の推移に応じてラベルに対応する文書集合の特徴や分布が大きく変化するコンセプトドリフトが発生する [4][42][43].

先行研究ではストリームを対象としたマルチラベル分類手法が提案されている．Xioufis らは，ラベルごとに当該のラベルに属する文書と属さない文書をそれぞれ別のウィンドウに保持することで，各ラベルの出現の割合に影響されず，直近に出現した文書の特徴も加味した分類手法を提案している [46]．ここでは，k-NN を用いてラベルごとに多重ウィンドウ内の文書でテスト文書と類似度が高い k 個の文書を抽出する．k 文書が各ウィンドウに含まれる数を計算し，閾値によってラベルを決定する．ストリーム中で学習データが到着するたびに，最も古い文書と到着した文書が入れ替えウィンドウを遷移させる．Read らは，Hoeffding Tree をマルチラベルに拡張した Multi-label Hoeffding Tree(MLHT) を提案している [29]．MLHT ではラベルの出現確率と非出現確率を用いたエントロピーによりノードを分割して決定することで決定木を構築する．文書ストリーム中で能動学習を用いた分類手法が提案されている [43]．Wang らは，新たに到着

した文書チャンク内で能動学習を行い，ラベル無し文書を適切に抽出し，ラベル有り文書と共に分類器を構築する手法を提案している．

トピックモデルは文書集合のモデル化手法として，様々な応用例が示されている．時系列データへの適用として，Blei らは Dynamic topic model(DTM) を提案し，時間単位のトピックの変化を捉えている [**?**]．Naveed らによる Author-Topic-Time(ATT) モデルは AT モデル [31] と TOT モデル [44] の発展形であり，学習データとしてタイムスタンプを用いる [27]．このモデルではトピックと著者にタイムスタンプによる時間分布を与え，時間によるトピックと著者の出現の変化を考慮している．文書及び時期によるトピック分布と著者の分布を捉える．ただし，新たに文書が発生する動的な文書集合を対象としていない．

トピックモデルはパラメータ学習を一括して行うため，大量の学習データに対しては極めて学習効率が悪い．この問題を解決するためにトピックモデルのオンライン学習では [7][32][14][50]，差分型ギブスサンプリングや EM アルゴリズムにより，一括学習の適用が困難な膨大な文書集合を効率よく扱う．また，ストリームデータなどの動的な文書集合に適用すれば，新たな特徴を学習することができる．Canini らはギブスサンプリングを基にした差分型サンプリング手法を提案している [7]．この手法では，現在までに得られたパラメータを基に各単語にトピックを割り当て，その後活性化ステップにおいて過去のトピック割り当てをランダムに選択し，再サンプリングを行う．Iwata らによる Online Multiscale Dynamic Topic モデルはオンライン学習可能な DTM の発展形であり，異なるスケールでの単語の依存性を用い，複数のスケールでトピックの時間発展を捉える [39]．本稿では異なる性質を持つトピックを扱う．Kawamae らは，タイムスタンプが与えられた文書集合に対して，定常トピックと時制トピックを検出する Theme Chronicle Model(TCM) を提案している [17]．TCM は，切り替え変数により，話題，トレンド，文書依存，バックグラウンドという 4 種類のトピックを学習する．話題は文書ごとにトレンドの分布を持ち，タイムスタンプによりトレンドの変化を扱う．しかし，ラベルごとのトピック分布やその変化を直接的には考慮していない．

8.3 文書ストリームのマルチラベル分類

8.3.1 問題設定

本研究で扱う文書ストリームは，ラベル有り文書列 $\mathbf{L} = \{l_1, l_2, ..., l_{|L|}\}$ とラベル無しの文書列 $\mathbf{U} = \{u_1, u_2, ..., u_{|U|}\}$ の混合から構成されるとする．文書分類では，初期の学習文書集合から分類器の特徴を学習し，ストリーム中で新たに到着するラベル無しの文書 u のラベルを推定する．また，到着した文書がラベル有り文書 l である場合は追加の学習を行う．

8.3.2 マルチラベル分類

マルチラベル分類は，文書が複数のラベルを有するとし，各文書に対してラベルベクトル $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ を推定する．ここで y_i は $y_i \in \{0, 1\}$ であり，属するラベルの値が 1，属さないラベルが 0 となる．ラベルベクトルを推定するため，McCallum ら [23] は，ガウス分布を前提とした EM アルゴリズムを用いる方式を提案している．ラベルの組合わせの個々に対して，各確率分布パラメータと混合比を計算し，最大尤度のマルチラベルを割り当てる．しかし，ラベルの組み合わせ数は $2^n - 1$ であるため，ラベル数の増加に伴い計算時間が膨大となる．このため，サンプリングや事後確率の収束を用いた近似方式がある [28]．しかし，共起性の高いラベルの組み合わせを考慮していない．文書ストリームではラベルの出現確率が大きく変動するため，このラベル間の関係の変化も扱う必要がある．

本稿は，マルチラベルの組み合わせをラベル有り文書がウィンドウ内で出現した組み合わせに限定し，ウィンドウ内でのマルチラベルの出現確率によりラベル間の共起関係を利用する．また，ストリーム中で到着するラベル有り文書から新たなラベルの組み合わせを獲得する．この手法により大幅な効率向上が期待できる．

8.4 トピックモデル

トピックモデルとは, 1 つの文書が複数のトピックの混合として表現されるという仮定である．1 つの文書が 1 つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある. その中でも最近用いられているのが Latent Dirichlet Allocation(LDA) である [5]．確率的潜在意味索引付け (Probabilistic Latent Semantic Indexing, pLSI) では, LDA と違って, トピックの混合比を学習データの文書集合に依存して固定化している [13]. 一方, LDA ではこの混合比を事前分布から生成する点で異なる．LDA のパラメータ推定では，潜在変数であるトピック z，文書ごとのトピックの確率分布 θ，トピックの単語分布 ϕ を文書集合に対して尤度が最大となるように推定する．この潜在変数の推定には，ギブスサンプリング等が用いられる．ギブスサンプリングでは，サンプリングにより 1 つの単語に対して 1 つのトピックを割り当てる．この割り当ては，更新を行う単語以外のすべてのトピックの割り当てによって更新される．

2 章図 2.1(a) では，LDA のグラフィカルモデルを示す. 図中の変数は, 図 2.1 左下に, ディリクレ事前分布 $Dir(\beta)$，図 2.1 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1 左上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N_d は各文書の単語数を表す．LDA の単語生成過程を以下で示す. まず, すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し, 同様に, すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する. 次に, 文書 d 内の i 番目の単語 w_i において，抽出した文書 d の多項分布 $Multinomial(\theta_d)$

からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する．
文書 d の単語 w_i のトピックが z_i となる確率は，以下の式で求まる．

$$
\begin{aligned}
&P(z_i|\mathbf{z}_{N\backslash i}, \mathbf{w}_N) \\
&\propto \frac{n^{w_i}_{z_i,N\backslash i} + \beta}{n^{(.)}_{z_i,N\backslash i} + V\beta} \frac{n^{d}_{z_i,N\backslash i} + \alpha}{n^{d}_{(.),N\backslash i} + T\alpha}
\end{aligned}
\tag{8.1}
$$

ここで，$\mathbf{z}_{N\backslash i} = \{z_1, ..., z_{i-1}, z_{i+1}, ..., z_N\}$ であり，i 番目の単語の割り当てを除外することを表す．$n^{w_i}_{z_i}$ は単語 w_i にトピック z_i が割り当てられた回数，$n^{(.)}_{z_i}$ は全単語中で z_i が割り当てられた合計である．$n^{d}_{z_i}$ は，文書 d で z_i が割り当てられた回数，$n^{d}_{(.)}$ は d にトピックが割り当てられた合計である．

LDA の拡張である AT モデルでは，2 章図 2.1(b) で表される変数 x により文書の著者などの該当の文書が属するラベルの情報を持つ．文書 d において $\mathbf{a}_d$ から著者 x は一様に得られ，トピック z は文書のトピック分布に基づきトピックから単語が生成される．LDA は文書ごとにトピック分布を持つが，AT モデルでは著者ごとにトピック分布を持つ点で異なる．また，AT モデルは文書に複数のラベルが存在することを仮定している．しかし，文書内でのラベルの偏りを考慮しておらず各ラベルは一様に生成される．提案モデルは AT モデルと比較して，変動トピック分布を取り入れる，文書内のラベルの出現確率の偏りを考慮するという 2 点で異なる．文書 d の単語 w_i の著者が x_i，トピックが z_i となる確率は，以下の式で求まる．

$$
\begin{aligned}
&P(z_i, x_i|\mathbf{z}_{N\backslash i}, \mathbf{w}_N, \mathbf{a}_d) \\
&\propto \frac{n^{w_i}_{z_i,N\backslash i} + \beta}{n^{(.)}_{z_i,N\backslash i} + V\beta} \frac{n^{x_i}_{z_i,N\backslash i} + \alpha}{n^{x_i}_{(.),N\backslash i} + T\alpha}
\end{aligned}
\tag{8.2}
$$

ここで，$n^{x_i}_{z_i}$ は z_i が著者 x_i に割り当てられた回数，$n^{x_i}_{(.)}$ は全トピックで x_i が割り当てられた合計である．

これらのモデルは一般的に静的な文書集合に対して適用される．LDA をストリームに対応させるためにオンライン学習手法が示されている [2]．Banerjee らによるオンライン学習では，まず現在までの文書集合に対して，パラメータを推定する．得られたパラメータを基に新たな文書のパラメータを MAP 推定する．新たな文書 d の単語 w_i のトピック割り当ては以下の式で推定する．

$$
\begin{aligned}
&P(z_i|\mathbf{z}_{i-1}, \mathbf{w}_i) \\
&\propto \frac{n^{w_i}_{z_i,i\backslash i} + \beta}{n^{(.)}_{z_i,i\backslash i} + V\beta} \frac{n^{d}_{z_i,i\backslash i} + \alpha}{n^{d}_{(.),i\backslash i} + T\alpha}
\end{aligned}
\tag{8.3}
$$

一括学習では単語 w_i のトピックを $\mathbf{z}_{i,N\backslash i}$ より推定し，単語列 $\mathbf{w}_N$ 内で繰り返し学習するが，オンライン学習では，現在までの観測データ $\mathbf{z}_{i-1}$ を用いて w_i のトピックを推定する．ここでは，過去の文書のトピックに関して再推定は行わない．

8.5 提案手法

本稿では，文書ストリーム $\mathbf{D} = \mathbf{L} \cup \mathbf{U}$ に対してトピックモデルを用いてマルチラベル分類を行う．Banerjee らによるオンライン学習と同様にまず初期の m 個の学習文書 $\{l_1, l_2, ..., l_m\}$ から，ラベルごとの定常トピック分布 θ^{st}，変動トピック分布 θ^{tr} と各トピックの単語分布 ϕ^{st}，ϕ^{tr} を推定する．新たに到着する文書 d_i が，ラベルを有する $d_i \in \mathbf{L}$ のとき追加の学習文書としてオンライン学習を適用する．または，d_i がラベル無し文書 $d_i \in \mathbf{U}$ のとき，提案モデルから得られたパラメータを用いて文書のラベルベクトル $\mathbf{y}_i$ を推定する．

提案モデルは AT モデルを拡張し，ストリーム中での特徴の変動を考慮するため新たなトピック分布を加え，単語分布の変化を表す．定常トピック分布はラベル内の全ての文書から学習されるが，変動トピック分布はウィンドウ内で出現頻度の高くなる単語を学習する．この 2 つのトピック分布を TCM と同様に指示変数を用いて単語ごとに切り替える [17]．この提案モデルにより，複数のラベルを持つ文書から各ラベルに対応する文書集合の特徴を抽出する．テスト文書のラベルの推定では，ウィンドウ内で共起したラベルに限定し，共起関係を考慮する．ウィンドウ S は最新のラベル有り文書 l_t から h 個とし，$S = \{l_t, l_{t-1}, ..., l_{t-h}\}$ と定義する．新たなラベル有り文書が到着するたびにウィンドウが遷移し，最も古いラベル有り文書と新たなラベル有り文書が入れ替わる．

8.5.1 提案モデル

本提案では，定常トピック分布及び変動トピック分布という 2 つの種類のトピックを用いる．現実の出来事に関連して短期間で話題が急激に変化するといったストリームの特性を加味したモデル化を行う．提案方式のグラフィカルモデルを図 8.1 に示す．図中の変数は表 8.1 に示す．

生成過程において，まず, ディリクレ事前分布 $Dir(\delta)$ より，文書に付与されたマルチラベル $\mathbf{a}_d$ に含まれる文書 d 内のラベルの確率分布である μ_d を抽出し，文書 d 内の i 番目の単語 w_i のラベル x_i を選択する．ついで，ベータ事前分布 $B(\gamma)$ から λ_d を抽出し，w_i が，定常トピックか変動トピックのどちらから生成されるかを選択する．その指示変数を k_i とする．次に選択されたトピックのディリクレ事前分布 $Dir(\alpha^{k_i})$ からラベルのトピック分布である $\theta_{x_i}^{k_i}$ を抽出する．多項分布 $Multinomial(\theta_{x_i}^{k_i})$ からトピック z_i を選択する．そのトピック z_i の多項分布 $Multinomial(\phi_{z_i}^{k_i})$ から単語 w_i を生成する．

表 8.1: グラフィカルモデルのパラメータ

α	ディリクレ分布のパラメータ
θ^{st}	定常トピックの確率
θ^{tr}	変動トピックの確率
z	トピック
w	単語
A	全ラベル数
β	ディリクレ分布のパラメータ
ϕ^{st}	定常トピックの単語の確率
ϕ^{tr}	変動トピックの単語の確率
N_d	d 番目の文書の単語数
D	文書数
T	トピック数
γ	ベータ分布のパラメータ
λ	トピック分布の確率
k	定常トピックか変動トピックの割り当て
$\mathbf{a}_d$	d 番目の文書のマルチラベル
δ	ディリクレ分布のパラメータ
μ	文書内のラベルの出現確率
x	ラベルの割り当て

定常トピック分布はすべての文書集合から学習されるが，変動トピック分布はウィンドウ S 内で局所的に学習される．

文書 d の単語 w_i のラベルが x_i，トピックが z_i となる確率は，ギブスサンプリングにより以下の式で近似される．

$$
\begin{aligned}
&P(z_i, x_i, k_i \in \{(st), (tr, S)\} | \mathbf{z}_{N\backslash i}, \mathbf{w}_N, \mathbf{a}_d) \\
&\propto \frac{n^{k_i, w_i}_{z_i, N\backslash i} + \beta^{k_i}}{n^{k_i, (.)}_{z_i, N\backslash i} + V\beta^{k_i}} \frac{n^{k_i}_{x_i, z_i, N\backslash i} + \alpha^{k_i}}{n^{k_i}_{x_i, (.), N\backslash i} + T\alpha^{k_i}} \\
&\frac{n^{k_i}_{d, N\backslash i} + \gamma}{n^{(.)}_{d, N\backslash i} + 2\gamma} \frac{n^{x_i}_{d, N\backslash i} + \delta}{n^{(.)}_{d, N\backslash i} + A_d\delta}
\end{aligned}
\tag{8.4}
$$

$$
\phi^{k,w}_z = \frac{n^{k,w}_z + \beta^k}{n^{k,(.)}_z + V\beta^k} \tag{8.5}
$$

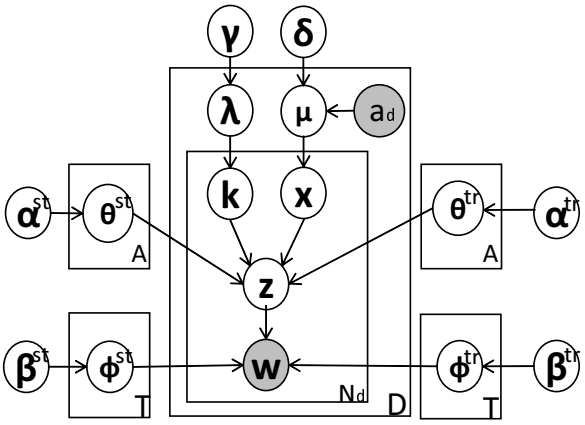

図 8.1: 提案モデル

$$\theta^k_{x,z} = \frac{n^k_{x,z} + \alpha^k}{n^k_{x,(.)} + T\alpha^k} \tag{8.6}$$

(st) は定常分布を表し，(tr,S) はウィンドウ S 内の変動分布を表す．k_i は指示変数であり単語 w_i に定常トピック st と変動トピック tr のどちらが割り当てられたかを表す．n^{k_i,w_i}_{zi}, $n^{k_i,(.)}_{z_i}$ は，トピック分布 k_i から w_i が z_i に割り当てられた回数，k_i から z_i に割り当てた回数の合計，$n^{k_i}_{x_i,z_i}$，$n^{k_i}_{x_i,(.)}$ はラベル x_i に k_i から z_i に割り当てられた回数，x_i に k_i から割り当てた回数の合計，$n^{(tr,S)}$ は変動トピック分布からウィンドウ S 内で割り当てられた回数を表す．$n^{k_i}_d$ は文書 d で k_i に割り当てられた回数, $n^{x_i}_d$ は d で x_i に割り当てられた回数を示す．V は全単語数，T は全トピック数，A_d は d のラベル数である．

ストリーム中では，新たに到着するラベル有り文書 $\{l_{m+1}, ..., l_{|L|}\}$ を用いて，逐次的にオンライン学習を行う．初期学習では m 個の学習文書内でギブスサンプリングにより繰り返し学習を行うのに対し，追加の学習では新たに到着した文書のパラメータのみ学習する．到着した文書 d の i 番目の単語のラベルとトピックは以下の式で求める．

$$\begin{aligned} &P(z_i, x_i, k_i \in \{(st), (tr, S)\} | \mathbf{z}_{i\backslash i}, \mathbf{w}_i, \mathbf{a}_d) \\ &\propto \frac{n^{k_i,w_i}_{z_i,i\backslash i} + \beta^{k_i}}{n^{k_i,(.)}_{z_i,i\backslash i} + V\beta^{k_i}} \frac{n^{k_i}_{x_i,z_i,i\backslash i} + \alpha^{k_i}}{n^{k_i}_{x_i,(.),i\backslash i} + T\alpha^{k_i}} \\ &\frac{n^{k_i}_{d,i\backslash i} + \gamma}{n^{(.)}_{d,i\backslash i} + 2\gamma} \frac{n^{x_i}_{d,i\backslash i} + \delta}{n^{(.)}_{d,i\backslash i} + A_d\delta} \end{aligned} \tag{8.7}$$

8.5.2 ラベリング手法

ラベリング手法では，テスト文書のマルチラベルの推定に，提案モデルにより学習されたパラメータとウィンドウ内でのマルチラベルの出現確率を用いる．マルチラベ

ルの出現確率は新たにラベル有り文書が到着することで逐次変化する．提案手法では，マルチラベルの組み合わせをウィンドウ内で出現したマルチラベルに限定する．また，推定結果となるマルチラベルは，各ラベルから生成される尤度が最大となるラベルを含むマルチラベルとする．本手法では，正解のマルチラベルがウィンドウ内に出現していない場合，正しいマルチラベルを推定できないが，共起性の高いマルチラベルを推定結果とすることを優先する．

まず各ラベルからテスト文書 d が生成される尤度を計算する．この結果から最も尤度の高い単一ラベル y_j を決定する．次いでウィンドウ内でその最尤ラベルを含むすべてのマルチラベルである候補マルチラベル $G = \{g_1, g_2, ..., g_{|G|}\}$ を抽出する．最後に候補マルチラベルの中で事後確率が最大となるマルチラベル g_c をラベルベクトルの推定結果とする．ラベリングの手順を以下に示す．

1. ラベル集合 $\{y_1, ..., y_n\}$ から尤度が最大となる y_j を求める．
2. ウィンドウ S 内でラベル y_j を含むすべてのマルチラベルを求める．
3. ラベルセットの中から事後確率が最大となるラベルを求める．

$P(d|y)$ はラベル y からテスト文書が生成される尤度であり，以下の式で求める．

$$P(d|y) = \prod_{w}^{N_d}[\sum_{z}^{T} \frac{n_z^{(st),w} + \beta^{st}}{n_z^{(st),(.)} + V\beta^{st}} \frac{n_{y,z}^{(st)} + \alpha^{st}}{n_{y,(.)}^{(st)} + T\alpha^{st}} + \sum_{z}^{T} \frac{n_z^{(tr,S),w} + \beta^{tr}}{n_z^{(tr,S),(.)} + V\beta^{tr}} \frac{n_{y,z}^{(tr,S)} + \alpha^{tr}}{n_{y,(.)}^{(tr,S)} + T\alpha^{tr}}] \tag{8.8}$$

推定結果となるマルチラベルは以下の式で求める．

$$Ans = arg \max_{g \in G} P(g)P(d|g) \tag{8.9}$$

ここで，$P(g)$ はマルチラベルの出現確率である．$P(d|g)$ はマルチラベルからテスト文書が生成される尤度であり，以下の式で求める．

$$P(d|g) = \prod_{w}^{N_d}[\sum_{z}^{T} \frac{n_z^{(st),w} + \beta^{st}}{n_z^{(st),(.)} + V\beta^{st}} \frac{\sum_{y \in g}^{|g|}(n_{y,z}^{(st)} + \alpha^{st})}{\sum_{y \in g}^{|g|}(n_{y,(.)}^{(st)} + T\alpha^{st})} + \sum_{z}^{T} \frac{n_z^{(tr,S),w} + \beta^{tr}}{n_z^{(tr,S),(.)} + V\beta^{tr}} \frac{\sum_{y \in g}^{|g|}(n_{y,z}^{(tr,S)} + \alpha^{tr})}{\sum_{y \in g}^{|g|}(n_{y,(.)}^{(tr,S)} + T\alpha^{tr})}] \tag{8.10}$$

$$P(g) = \frac{n_g^S}{h} \tag{8.11}$$

ここで n_g^S はウィンドウ S 内で，マルチラベル g が出現した回数である．h はウィンドウの文書数である．

表 8.2: 実験データ

	C15	C151	C1511	C152	C18	C181	CCAT	E21	E212	ECAT
学習文書数	828	394	105	465	258	268	1064	139	139	142
テスト文書数	15206	7878	2627	7853	4146	4241	18918	2921	2914	2834
	GCAT	GSPO	M11	M12	M13	M131	M14	M141	M143	MCAT
学習文書数	546	307	396	187	253	254	470	268	168	1216
テスト文書数	9163	5573	6346	3422	3496	3478	8462	5271	2622	20719

8.6 実験

実験では Reuters Corpus(RCV1) を用いてマルチラベル分類を行う．比較手法には，他のストリーム中のマルチラベル分類手法との比較として多重ウィンドウ (MW) による分類手法 [46] を用いる．トピックモデルの比較として提案手法の基である AT モデル [31] を用いる．単語の確率分布を使用する手法としてマルチラベル分類に対応した単純ベイズ (MLNB)[52] を用いる．AT モデルでは，提案手法と同様にオンライン学習を行い，提案ラベリング手法を用いる．AT モデルは変動分布を用いないため，変動分布の計算は除く．MLNB はストリームに対する分類手法ではないため，MLNB では新たな学習文書が到着するたびに一括学習により再学習を行う．

8.6.1 実験準備

実験に用いる Reuters Corpus は 1996 年 8 月 20 日から 1997 年 8 月 19 日までの 1 年分のニュース記事であり，1 つの記事に 128 種類からなるラベルが複数付いている．Reuter Corpus はニュース記事であることから，該当のラベル内で繰り返し使用される固有名詞やそれに関連した単語などの定常的な特徴を有する．また，ある時刻に起こった出来事に関する話題や新たに発生した話題に関連する単語など変動的な特徴を含む． 本実験では，頻度が上位となるマルチラベルに含まれるラベルを頻度が高い順に 20 個抽出して用いる．用いるラベルと各ラベルを含む文書数を表 8.2 に示す．表 8.3 は各ラベルの識別子である．実験データには各月の先頭から 5 日間で計 60 日分を用いる．総文書数は 54676 文書であり 3000 文書を初期の学習文書として，残りの 51676 文書をテスト文書とする．テスト文書中の 10%を等間隔に到着する新たなラベル有り文書とし，分類後に文書のラベルを使用して学習に用いる．ラベル有り文書の到着間隔が精度に影響する可能性があるため，5000 文書ごとに 10%である 500 文書がランダムに到着する場合と分類精度を比較する．各文書には平均 2.67 個のラベルが付与されている．実験データ中の不要語は取り除く．

提案手法の各パラメータの値はトピック数 175，ギブスサンプリングの繰り返し回数 200 回とし，事前分布のパラメータは α^{st}=0.5，α^{tr}=0.5，β^{st}=1/語彙数，β^{tr}=1/語彙数，γ=0.1，δ=0.1 とする．また，ウィンドウの文書数は 3000 とする．AT モデルではトピック数を 200 とする．

表 8.3: ラベル識別子

ラベル名	種類
C15	PERFORMANCE
C151	ACCOUNTS/EARNINGS
C1511	ANNUAL RESULTS
C152	COMMENT/FORECASTS
C18	OWNERSHIP CHANGES
C181	MERGERS/ACQUISITIONS
CCAT	CORPORATE/INDUSTRIAL
E21	GOVERNMENT FINANCE
E212	GOVERNMENT BORROWING
ECAT	ECONOMICS
GCAT	GOVERNMENT/SOCIAL
GSPO	SPORTS
M11	EQUITY MARKETS
M12	BOND MARKETS
M13	MONEY MARKETS
M131	INTERBANK MARKETS
M14	COMMODITY MARKETS
M141	SOFT COMMODITIES
M143	ENERGY MARKETS
MCAT	MARKETS

8.6.2 評価方法

実験の評価には全体の分類精度を比較するためにf値を用いる．また，提案手法では最尤となるラベルを基にマルチラベルを推定するため，最尤となるラベルが正解に含まれているかどうかが全体の精度に大きな影響を及ぼす．これを評価するために予測ラベル中で最も尤度の高いラベルのエラー率を表すOne-Errorを用いる．f値は再現率と適合率の調和平均であり，再現率は実際に正であるもののうち，正であると予測されたものの割合，適合率は正と予測したデータのうち，実際に正であるものの割合である．各ラベルのf値を以下の式で求める．

$$R_i = \frac{a_i}{a_i + c_i} \tag{8.12}$$

$$P_i = \frac{a_i}{a_i + b_i} \tag{8.13}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が異なった数である．この2つの式の調和平均であるf値を次のように定義する．

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}$$

また，マルチラベルのf値を正解のラベルベクトル $y_i = (y_{i,1}, ..., y_{i,n})$ と予測したラベルベクトル $y_i^{'} = (y_{i,1}^{'}, ..., y_{i,n}^{'})$ から以下の式で求める．

$$multiR_i = \frac{\sum_{l=1}^{L} y_{n,l} y_{n,l}^{'}}{\sum_{l=1}^{L} y_{n,l}} \tag{8.14}$$

$$multiP_i = \frac{\sum_{l=1}^{L} y_{n,l} y_{n,l}^{'}}{\sum_{l=1}^{L} y_{n,l}^{'}} \tag{8.15}$$

$$multi - f_i = \frac{2P_i R_i}{P_i + Ri} \tag{8.16}$$

One-Error は以下の式で求める．

$$one - error(f) = \frac{1}{P} \sum_{i=1}^{P} \Big((arg \ \max_{y \in Y} f(x_i, y)) \notin Y_i \Big) \tag{8.17}$$

また，トピックが持つ単語分布の変化を比較するのに JS ダイバージェンスを用いる．JS ダイバージェンスは対称な確率分布の差の尺度であり確率分布 P と確率分布 Q が与えられたとき，JS ダイバージェンスは以下の式で求まる．

$$JS(P//Q) = \frac{1}{2}(\sum_{x} P(x) log \frac{P(x)}{R_(x)}$$
$$+ \sum_{x} Q(x) log \frac{Q(x)}{R(x)}) \tag{8.18}$$

ここで R は確率分布 P と Q の平均であり，$R = \frac{P+Q}{2}$ となる．JS ダイバージェンスの値は P=Q のとき 0 となり，確率分布の差が大きいほど値が大きくなる．JS ダイバージェンスにより，テスト文書中で 10000 文書ごとに前後の時点とのトピックの単語分布を比較し，すべてのトピックで合計する．これにより，各時点でのトピックの中身の変化を測る．

8.6.3 実験結果

実験結果を表 8.4，表 8.5，表 8.6，表 8.7，表 8.8 に示す．表 8.6 より，ラベルごとの全体の f 値は提案手法で 0.793，AT で 0.747，MW で 0.746，MLNB で 0.571 となっており，提案手法の f 値が最も高い．また，表 8.4，表 8.5 より再現率と適合率においても提案手法が最も高い値を示している．各ラベルの f 値を比較すると 20 ラベル中 14 のラベルで提案手法の f 値が最も高い．MLNB では再現率が最も高いラベルが多いが，適合率が他の手法と比較して 0.3 程度低いため，全体の f 値は 4 つの手法中で最も悪くなっている．表 8.7 より，等間隔にラベル有り文書が到着する場合では，multi-f 値は提案手法で 0.842，AT で 0.785，MW で 0.795，MLNB で 0.685 となっている．提案手法は他の手法と比較して multi-f 値が 4.7%から 15.7%上昇している．ランダムにラベル有り文書が到着する場合では，multi-f 値は提案手法で 0.842，AT で 0.781，MW で 0.792，MLNB で 0.680 となっている．到着間隔を変化させた場合においても提案手法が最も高い精度を示している．表 8.8 より，One-Error は提案手法で 0.074，AT で 0.174，MW で 0.132，MLNB で 0.163 であり，提案手法が最もエラー率が低い．

表 8.4: 再現率

	提案手法	AT	MW	MLNB
C15	0.726	0.658	**0.772**	0.748
C151	**0.781**	0.535	0.735	0.756
C1511	0.539	0.606	0.585	**0.722**
C152	**0.764**	0.736	0.524	0.615
C18	0.708	0.660	0.584	**0.795**
C181	0.722	0.745	0.601	**0.807**
CCAT	0.821	0.803	**0.856**	0.844
E21	0.815	0.814	0.762	**0.858**
E212	0.817	0.816	0.763	**0.860**
ECAT	0.755	0.817	0.762	**0.838**
GCAT	**0.979**	0.953	0.886	0.923
GSPO	**0.979**	0.926	0.900	0.976
M11	**0.865**	0.844	0.687	0.605
M12	**0.726**	0.597	0.442	0.656
M13	0.766	0.664	0.703	**0.855**
M131	**0.915**	0.769	0.699	0.840
M14	0.877	0.724	0.857	**0.927**
M141	**0.930**	0.853	0.720	0.891
M143	0.951	0.784	0.849	**0.953**
MCAT	**0.876**	0.713	0.869	0.834
全体	**0.816**	0.751	0.728	0.815

8.7 考察

本手法は比較手法に用いた MW，AT モデル，MLNB と比較して最も高い精度を示している．トピックモデルの一種である本手法と AT モデルの比較では，提案手法が f 値で 8.6%高くなっている．提案手法では，ウィンドウ内に限定した変動トピック分布を用いることで，変動トピックを用いない AT モデルと比較して精度が上昇している．表 8.12 より，提案手法の定常トピックと AT モデルでは学習文書が増えるにつれてトピック内の変動が単調に減少している．変動トピック分布は，JS ダイバージェンスが単調に減少せずに増減しているためトピックの中身が変化している．

表 8.9 より，トピック数を変化させた場合，全てのトピック数で提案手法の精度が高い．これは，提案手法では定常トピック分布と変動トピック分布の 2 つを持つことによる．1 種類のトピックしか用いない AT モデルは，トピック数が増えても提案手法と

表 8.5: 適合率

	提案手法	AT	MW	MLNB
C15	0.757	0.825	**0.835**	0.779
C151	0.821	**0.875**	0.852	0.611
C1511	0.411	0.389	**0.502**	0.091
C152	0.586	0.559	**0.682**	0.479
C18	**0.686**	0.567	0.674	0.246
C181	**0.704**	0.604	0.680	0.252
CCAT	0.789	0.806	**0.881**	0.861
E21	0.811	0.734	**0.828**	0.187
E212	0.811	0.734	**0.828**	0.187
ECAT	0.733	0.677	**0.816**	0.218
GCAT	0.905	**0.947**	0.908	0.832
GSPO	**0.986**	0.973	0.886	0.611
M11	0.664	0.580	**0.717**	0.480
M12	**0.763**	0.632	0.567	0.268
M13	0.692	**0.774**	0.709	0.210
M131	0.720	**0.784**	0.710	0.212
M14	**0.948**	0.910	0.886	0.716
M141	**0.890**	0.835	0.766	0.420
M143	**0.896**	0.814	0.686	0.242
MCAT	0.851	0.835	0.872	**0.891**
全体	**0.771**	0.743	0.764	0.440

同等の精度を達成できていない.

表 8.7 より，到着間隔を変化させることで multi-f 値が若干変化しているが，全体の傾向は変わっていない．テスト有り文書の到着間隔は全体の精度に影響を及ぼす可能性があるが，提案手法は等間隔に到着する場合に限らず有効であると考えられる.

表 8.10 より，累積の multi-f 値のストリーム内の推移は，10000 文書目でパラメータの更新を行うとき 0.829，更新無し 0.821，50000 文書目で提案手法 0.842，更新無し 0.806 となる．本手法では multi-f 値が 1.3%上昇しているが，更新無しでは 1.5%減少している．更新を行わなかった場合，テスト文書の特徴が初期の学習文書の特徴と異なっていくため分類精度が低下する．本手法では新たな特徴を学習しているため精度が上昇している.

表 8.11 より，各区間の multi-f 値は，全ての区間において提案手法が最も高い．提案手法の multi-f 値は 0〜5000 の区間で 0.830，45000〜50000 の区間で 0.838 と分類を継

表 8.6: f 値

	提案手法	AT	MW	MLNB
C15	0.741	0.732	**0.802**	0.763
C151	**0.801**	0.664	0.789	0.676
C1511	0.466	0.474	**0.540**	0.162
C152	**0.663**	0.635	0.593	0.539
C18	**0.697**	0.610	0.626	0.375
C181	**0.713**	0.667	0.638	0.385
CCAT	0.805	0.805	**0.869**	0.852
E21	**0.813**	0.772	0.794	0.307
E212	**0.814**	0.773	0.794	0.307
ECAT	0.744	0.741	**0.789**	0.346
GCAT	0.941	**0.950**	0.897	0.875
GSPO	**0.982**	0.949	0.893	0.752
M11	**0.752**	0.688	0.702	0.535
M12	**0.744**	0.614	0.496	0.381
M13	**0.727**	0.715	0.706	0.337
M131	**0.806**	0.777	0.704	0.338
M14	**0.911**	0.806	0.872	0.808
M141	**0.909**	0.844	0.742	0.571
M143	**0.923**	0.799	0.759	0.386
MCAT	0.863	0.769	**0.870**	0.862
全体	**0.793**	0.747	0.746	0.571

続しても精度を維持している．MLNB は，再学習に一括学習を用いており，初期区間である 0〜5000 の区間で 0.547 と特に低いことから最も上昇率が高くなっている．しかし，35000〜40000 の区間と 40000〜50000 の区間を比較して+1.7%，続いて最後の区間と比較して+0.6%と学習文書が増えるにつれて上昇率は鈍化している．最も精度の高い最後の区間においても提案手法と比較して-9.1%と大きな差がある．

提案ラベリング法では，単一ラベルで最尤となったラベルを基に候補マルチラベルを抽出するため，最尤ラベルの推定が全体に大きな影響を与える．表 8.8 より，最尤ラベルの分類精度は非常に高い．また，表 8.13 より，マルチラベリング時の候補マルチラベル数別の multi-f 値は，0.720 から 0.905 となっている．候補マルチラベルの数が少ない 1 から 6 個の場合に 0.860 から 1 と特に高い精度になっている．候補マルチラベルの数が最大である 22 個の場合でも multi-f 値は 0.813 であり，候補数が多い場合においても高い精度を維持している．

表 8.7: multi-f 値

	提案手法	AT	MW	MLNB
(等間隔)				
multi-R	**0.850**	0.773	0.780	0.820
multi-P	**0.835**	0.798	0.811	0.588
multi-f	**0.842**	0.785	0.795	0.685
(ランダム)				
multi-R	**0.851**	0.776	0.778	0.817
multi-P	**0.832**	0.787	0.806	0.582
multi-f	**0.842**	0.781	0.792	0.680

表 8.8: OneError

	提案手法	AT	MW	MLNB
OneError	**0.074**	0.174	0.132	0.163

本稿では，ウィンドウ内の変動トピック分布とマルチラベルの出現確率を用いることで，multi-f 値が 0.842 と最も高精度になる．提案手法では動的学習により新たな特徴を学習し，ストリーム内のテスト文書の変化に影響されず分類精度を維持しており，文書ストリームの分類に有効である．

8.8 結び

本研究では定常分布と変動分布を考慮したトピックモデルとウィンドウ内で出現したマルチラベルの出現確率を用いたマルチラベル分類手法を提案した．本手法は，変動分布を用いることで，ストリーム中の話題の変化に影響されず各区間で高精度にマルチラベル分類が行える．マルチラベル分類の結果，multi-f 値は 0.842 と高精度で行えることを示した．また，新たに到着した学習データを用いてパラメータを更新することで精度が上昇することを示した．これにより提案手法を用いることでストリーム中のマルチラベル分類を高精度に行えることを示した．

表 8.9: トピック数別の multi-f 値

トピック数	100	125	150	175	200	250
提案手法	0.830	0.835	0.832	0.842	0.841	0.812
AT	0.779	0.779	0.766	0.777	0.785	0.784

表 8.10: 累積の multi-f 値の推移

テスト文書数	10000	20000	30000	40000	50000	difference
提案手法	0.829	0.838	0.840	0.842	0.842	+0.013 (+1.3%)
提案手法 (更新無し)	0.821	0.820	0.816	0.811	0.806	-0.015 (-1.5%)

表 8.11: 各区間の multi-f 値

テスト文書の区間	提案手法	AT	MW	MLNB
0〜5000	0.830	0.768	0.762	0.547
5000〜10000	0.829	0.776	0.822	0.589
10000〜15000	0.847	0.796	0.794	0.624
15000〜20000	0.848	0.791	0.794	0.657
20000〜25000	0.847	0.783	0.800	0.691
25000〜30000	0.839	0.780	0.776	0.706
30000〜35000	0.853	0.790	0.816	0.728
35000〜40000	0.844	0.784	0.792	0.724
40000〜45000	0.850	0.798	0.787	0.741
45000〜50000	0.838	0.780	0.802	0.747

表 8.12: JS ダイバージェンス

文書番号	0 ↔ 10000	10000 ↔ 20000	20000 ↔30000	30000↔ 40000	40000↔ 50000
定常トピック	11.05	7.82	5.26	4.20	3.69
変動トピック	16.48	11.93	10.22	10.31	10.43
AT	16.94	9.75	6.35	4.87	4.16

表 8.13: 最尤ラベルを含むマルチラベル数別の multi-f 値

候補ラベルセットの数	件数	multi-f 値
1	1718	1
2	1803	0.860
3	5977	0.885
4	3522	0.865
5	5693	0.831
6	1328	0.905
7	1805	0.789
8	7644	0.799
9	1700	0.767
10	3529	0.744
11	659	0.767
12	583	0.753
13	376	0.720
14	98	0.765
15	3680	0.732
16	2513	0.746
17	2779	0.783
18	2572	0.772
19	2211	0.790
20	135	0.812
21	343	0.813
22	1008	0.813

第9章　結論

9.1　本研究の貢献と効果

本研究では，潜在要因を考慮した確率モデルに基づき自然言語文書からの知識抽出を行う手法を提案した．自然言語文書からの知識抽出では，主に2つの問題について論じた．すなわち，文書を特徴付ける複数の要因が混在した文書集合からの特徴抽出と特徴の変化が起きる文書ストリームからの特徴抽出である．第1の問題では，トピックモデルにより単語の持つ潜在トピックを推定し，文書を潜在トピックの混合で表した．この潜在トピックを文書を特徴付ける要因と見なすことで，検索語の意味の抽出といった自然言語文書からの知識抽出が行えることを示した．また，日本語文書の品詞分布がジャンルを特徴付ける要因となることを示した．第2の問題では，ストリーム中での文書の特徴を抽出するためにトピックモデルのオンライン学習を行った．文書ストリーム中での特徴の変動を事前分布の変化と対応させることで特徴抽出を行った．また，ストリーム中での特徴を定常分布と変動分布として抽出した．

本研究では，これらの問題を網羅的に扱い，各問題を解決するための手法を提案した．第3章では，潜在トピックによる著者の特徴を抽出する手法について論じた．トピックモデルを用いて単語の潜在状態を推定することで文書をトピックの混合で表現した．これにより，クラスや話題といった文書を特徴付ける要因が混在した文書集合において高精度に分類が行えることを示した．

第4章では，日本語の品詞分布特性によるジャンルの特徴を抽出する手法について論じた．ここでは，日本語文書の品詞分布の特性をジャンルごとに仮定することで，文書の品詞情報のみを用いてジャンル分類を行えることを示した．これにより，品詞分布が文書を特徴付けていることを示した．

第5章では，検索語の意味を抽出する手法について論じた．検索語間の依存関係を捉えることで，検索語の意味を考慮して文書検索を行えることを示した．検索語に対して係り受け語を推定し，その組に対してトピックを推定することで，検索語の意味を抽出した．

第6章では，事前分布の学習による動的な特徴を抽出する手法について論じた．ストリーム中でのクラスの出現確率やトピックの確率分布，トピックの単語分布の変化を事前分布の変化として捉えることで，変動する特徴の抽出を行った．

第7章では，オンライン学習によるストリーム中の特徴を抽出するについて論じた．ストリーム中でオンライン学習を行うことで，文書集合の特徴の変化を抽出した．ストリーム中での定常的な特徴を定常トピック，変動的な特徴を変動トピックを用いて

表現することで文書ストリームの分類を行えることを示した．

第 8 章では，ストリーム中の複数のラベルを持つ文書からの特徴を抽出する手法について論じた．ここでは，文書中のラベルの混合と話題の混合を同時に扱えるようにトピックモデルを拡張した. ストリーム中でラベル間の共起関係を抽出することで，複数のラベルを持つ文書の分類を行えることを示した．

これらの手法を用いることで，自然言語文書からの知識抽出を行う上での各問題を解決することができる．

9.2 本研究の展開

本研究で提案した潜在要因を考慮した自然言語文書からの知識抽出技術は，文書の内容解析や情報選択の基盤技術としての利用が期待できる．例えば，感性情報を潜在要因として考慮したアンケート分析や商品のレビュー分析への応用が考えられる．また，提案手法は高精度に文書分類を行えることから，スパムメール識別や情報統合に適用可能である．オンライン学習を用いて確率モデルを更新しているため，毎日新規のレビューが発生するといった文書集合の動的な変化にも対応可能である．

9.3 本研究の発展

本研究では，確率モデルを用いて文書集合のモデル化を行った．この確率モデルは，現在の集合に対して構築したモデルであるため，集合の特徴の変化には自動的に対応できない．このため，文書ストリーム中で新たな特徴を捉えるために，オンライン学習により確率モデルを更新した．しかし，追加のラベル付き文書が学習に有効な文書であるか考慮していない．現実的には，Twitter でのハッシュタグ付きの文書といった新たなラベル付き文書は不定期に出現し，多くのノイズを含む．このため，確率モデルを更新するには多くのデータの中から学習に適切な文書だけを抽出する必要がある．ここでの問題は，大量の文書の中から人手によりどの文書が学習に有効か判別することは不可能である点にある．このため，コンピュータがモデルを修正するために適切な文書を自動的に選択することが望ましい．近年このような問題は能動学習として扱われている．能動学習は，大きく分けてプールベースとストリームベースに分けられる．プールベースの能動学習は，集合内のデータをランク付けし，学習に有効な文書を抽出する．ストリームベースの能動学習では，任意の基準を用いてデータが到着するたびに学習に用いるか判断する．能動学習は，コンピュータ自身が学習に有効な文書を検出できるという点で有用である．しかし，一般的な能動学習では，検出されたデータを人手によりラベル付けする必要がある．この過程はストリームを扱ううえで好ましくない．能動学習を新たに出現するラベル付きデータのみに適用すれば，人手によらずモデルを更新できる．しかし，学習データの総量が減少するため，性能を改

善できるかは定かではない．文書集合の変化に応じて自動的にモデルを更新し，頑強なモデルを構築することが今後の課題である．

また，学習した知識の利用に関して，本研究は確率モデルを学習した集合と抽出した知識を適用する集合が同一の特徴を持つ集合であることを仮定している．このため，集合内に存在するスポーツやITといったクラスが変化することを扱えない．転移学習は，情報源領域から得られた知識を異なる領域の解析に適用する枠組みである．学習データとテストデータが同一の領域であることを仮定しないため適用範囲が広く，様々な問題に転移学習が用いられている．ここでは，得られた知識を利用するために情報源領域と対象領域が関連していることが必要になる．しかし，対象領域に対する適切な情報源領域が存在するとは限らない．情報源領域と対象領域の関連性が低い場合，転移学習の精度が悪化することが知られている．このため，中間領域を用いて関連性の低い情報源領域と対象領域を繋ぐ転移学習手法が提案されている．今後は，確率モデルにより抽出した知識を如何に有効活用するかという問題について取り組む．

参考文献

[1] AlSumait, L., Barbara, D., Domeniconi, C.: On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking, proc. ICDM, pp.3-12, 2008

[2] Banerjee, A., Basu, S.: Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning, Proc. 7th SIAM International Conference on Data Mining, pp.437-442, 2007

[3] Berger, A. and Lafferty, J.: Information Retrieval as Statistical Translation, proc. ACM SIGIR, pp.222-229, 1999

[4] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., and Gavalda, R.: New ensenble methods for evolving data streams. In KDD '09, pp.139-148, 2009

[5] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp.993-1022, 2003

[6] Blei, D.M., and Lafferty, J.D.: Dynamic topic models, proc. ICML, pp.113-120, 2006

[7] Canini, K.R., Shi, L., Griffiths, T.L.: Online Inference of Topics with Latent Dirichlet Allocation, Journal of Machine Learning Research, Proceedings Track 5, pp.65-72, 2009

[8] Chai, K.M., Ng, H.T. and Chieu, H.L.: Bayesian Online Classifiers for Text Classification and Filtering, Proc.ACM (SIGIR), pp.97-104, 2002

[9] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R: Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6), pp.391-407, 1990

[10] Griffiths, T.L. and Steyvers, M.: Finding Scientific Topics, proc. National Academy of Sciences 101, suppl 1:5228-5235, 2004

[11] Grossman, D. A., Frieder, O.: Information Retrieval Algorithms and Heuristics (2nd ed.), pp.271-274, 2004, Springer Verlag

[12] Manku, G. S., Motwani, R.: Approximate Frequency Counts over Data Stream, Proc. 28th international conference on Very Large Data Bases (VLDB), pp.346-357, 2002

[13] Hofmann, T.: Probabilistic Latent Semantic Indexing, SIGIR, pp.50-57, 1999

[14] Hoffman, M.D., Blei, D.M., Bach, F.R.: Online Learning for Latent Dirichlet Allocation, proc. 24th Annual Conference on Neural Information Processing Systems (NIPS), pp.856-864, 2010

[15] Iwata, T., Yamada, T., Sakurai, Y., Ueda, N.: Online Multiscale Dynamic Topic Models, KDD'10, pp.663-672, 2010

[16] Holmes, D. and Forsyth, R.: The Federalist revised: New directions in authorship attribution, Literary and Linguistic Computing 10-2, pp.111-127, 1995

[17] Kawamae, N.: Theme chronicle model: chronicle consists of timestamp and topical words over each theme, Proc. 21st CIKM, pp. 2065-2069, 2012

[18] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, Proc. 8th SIGKDD, pp.91-101, 2002

[19] Kurohashi, S. and Nagao, M: KN Parser : Japanese Dependency/Case Structure Analyzer, proc. Workshop on Sharable Natural Language Resources, pp.48-55, 1994

[20] Li, W., McCallum, A.: Pachinko Allocation: DAG-structured mixture models of topic correlations, proc. ICML, pp.577-584, 2006

[21] Liu, X. and Croft, W. B.: Cluster-based retrieval using language models, proc. ACM SIGIR, pp.186-193, 2004

[22] Manning, C.D. and Schutze, H.: Foundations of Statistical Natural Language Processing, 1999, MIT Press

[23] McCallum, A.K.: Multi-Label Text Classification with a Mixture Model Trained by EM,Workshop on Text Learnong, AAAI '99 workshop on text learning, pp.1-7, 1999

[24] Mendelhall,T.C.: Mechanical Solution of a LiteraryProblem, Popular Science Monthly 60-2, 1901

[25] Minka, T. P.: Estimating a Dirichlet distribution, Technical report, M.I.T, 2000 (revised 2012)

[26] Nakayama, M. and Miura, T.: Identifying Topics by using Word Distribution, proc. PACRIM, 2007

[27] Naveed, N., Sizov, S., Staab, S.: ATT: Analyzing Temporal Dynamics of Topics and Authors in Social Media In: Proceedings of the ACM WebSci'11, pp. 1-7,2011

[28] Ramage, D., Hall, D., Nallapati, R. and Manning C.D.: Labeled LDA – a supervised topic model for credit attribution in multi-labeled corpora, proc. Empirical Methods in Natural Language Processing (EMNLP09), 2009, pp. 248-256

[29] Read, J., Bifet, A., Holmes, G. Pfahringer, B.,: Scalable and Efficient Multi-label Classification for Evolving Data Streams, Machine Learning, Volume 88, Issue 1-2, pp 243-272, 2012

[30] Lewis, D.D., et al.: RCV1 (Reuters Corpus Volume 1), 2004, `www.daviddlewis.com/resources/testcollections/rcv1/`

[31] Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence, pp.487-494, 2004.

[32] Sato, I., Kurihara, K., Nakagawa, H.: Deterministic Single-Pass Algorithm for LDA, proc. 24th Annual Conference on Neural Information Processing Systems (NIPS), 2074-2082, 2010

[33] Shinzato, K. and Kurohashi, S.: Exploiting Term Importance Categories and Dependency Relations for Natural Language Search, proc. 2nd Workshop on NLPIX 2010, pp.2-11, Beijing, China, 2010

[34] Shirai, M. and Miura, T.: On Domain Independence of Author Identification,IDEAL'11, pp.9-16, 2011

[35] Shirai, M. and Miura, T.: Document Classification Using POS Distribution, 16th East-European Conference on Advances in Databases and Information Systems(ADBIS), pp.346-356, 2012

[36] Shirai, M., Yanagisawa, T., Miura, T.: Context-based Query using Dependency Structures based on Latent Topic Model, JoDS Vol. 3, Issue 3, pp.157-168, 2014

[37] Shirai, M., Miura, T.: Online Classification with Partially Labelled Texts, International Conference on Web Intelligence (WI), pp.421-428, 2014

[38] Ida, Y., Nakamura, T., Matsumoto, T.: Domain-dependent/independent topic switching model for online reviews with numerical ratings, Proc. 22nd CIKM, pp. 229-238, 2013

[39] Iwata, T., Yamada, T., Sakurai, Y., Ueda, N.: Online multiscale dynamic topic models, proc. ACM SIGKDD, pp.663-672, 2010

[40] Wakabayashi,K. and Miura, T.: Forward-Backward Activation Algorithm for Hierarchical Hidden Markov Models, proc. 26th Annual Conference on Neural Information Processing Systems (NIPS), pp.1493-1501, 2012

[41] Wang, C., Blei, D., Heckerman, D.: Continuous Time Dynamic Topic Models, In UAI'08, pp.579-586, 2008

[42] Wang, H., Yin, J., Pei, J., Yu, P. S., and Yu, J. X.: Suppressing Model Overfitting in Mining Concept-Drifting Data Streams, In KDD '06, pp.736-741, 2006

[43] Wang, P., Zhang, P., and Guo, L.: Mining Multi-label Data Streams Using Ensemble-based Active Learning, SDM, pp 1131-1140, SIAM/Omnipres, 2012

[44] Wang, X., McCallum, A.: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, Proc. 12th SIGKDD, pp.424-433, 2006

[45] Wei, X. and W. Bruce Croft: LDA-Based Document Models for Ad-Hoc Retrieval, proc. ACM SIGIR, pp.178-185, 2006

[46] Xioufis, E. S., Spiliopoulou, M., Tsoumakas, G., Vlahavas, I.: Dealing with Concept Drift and Class Imbalance in Multi-Label Stream Classification, Proc. 22nd IJCAI, pp.1583-1588, 2011

[47] Yao, L., Mimno D., McCallum, A.: Efficient Methods for Topic Model Inference on Streaming Document Collections, In 15th ACM SIGKDD, pp.937-946, 2009

[48] Yanagisawa, T. and Miura, T. : Sentence Generation for Stream Announcement, proc. IEEE International Pacific Rim Conference on Communications, Computers and Signal Processing(PACRIM), 2009

[49] Yanagisawa, T., Miura, T. and Shioya, I.: Sentences generation by frequent parsing patterns, proc. 11th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), pp.53-62, 2010

[50] Yao, L., Mimno D., McCallum, A.: Efficient Methods for Topic Model Inference on Streaming Document Collections, proc.ACM SIGKDD international conference on Knowledge discovery and data mining, pp.937-946, 2009

[51] Yi, X. and Allan, J.: A Comparative Study of Utilizing Topic Models for Information Retrieval, proc. 31th European Conference on IR Research on Advances in Information Retrieval, pp.29-41, 2009

[52] Zhang, M. L., Pena, J. M., and Robles, V.: Feature Selection for Mining-Label Naive Bayes Classification, Information Sciences: an International Journal, Volume 179 Issue19, pp. 3218-3229, 2009

[53] Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad-hoc Information Retrieval, proc. ACM SIGIR, pp. 334-342, 2001

[54] 大野 晋: 基本語彙に関する二三の研究国語学,vol24,pp.34-46,1956

[55] 樺島　忠夫:　現代文における品詞の比率とその増減の要因について国語学,vol18,pp.15-20,1954

[56] 樺島 忠夫: 類別した品詞の比率に見られる規則性国語国文,vol250,pp.385-387,1955

[57] 金明哲: 読点と書き手の個性, 計量国語, Vol18,No8,pp.382-391,1993

[58] 金 明哲, 村上 征勝: ランダムフォレスト法による文書の書き手の同定, 統計数理,vol55,pp255-268,2007

[59] 松浦司，金田康正: n-gram 分布を用いた近代日本語小説文の著者推定, 自然言語処理, 134-5,1999

[60] 水谷 静夫: 大野の語彙法則について計量国語学,vol35,pp.1-13,1965